Chapter 6
MEDICAL CONCEPT REPRESENTATION

Christopher G. Chute

Mayo Clinic College of Medicine, Rochester, MN 55905

Chapter Overview

The description of concepts in the biomedical domain spans levels of precision, complexity, implicit knowledge, and breadth of application that makes the knowledge representation problem more challenging than that in virtually any other domain. This chapter reviews some of this breadth in the form of use-cases, and highlights some of the challenges confronted, including variability among the properties of terminologies, classifications, and ontologies. Special challenges arise at the semantic boundary between information and terminology models, which are not resolvable on one side of either boundary. The problems of aggregation are considered, together with the requirement for rule-based logic when mapping information described using detailed terminologies to high-level classifications. Finally, the challenge of semantic interoperability, arguably the goal of all standards efforts, is explored with respect to medical concept representation.

Keywords

vocabulary; ontology; classification; biomedical concepts; terminology

> ‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean – neither more nor less.’
> ‘The question is,’ said Alice, ‘whether you CAN make words mean so many different things.’
> ‘The question is,’ said Humpty Dumpty, ‘which is to be master – that's all.’
>
> Lewis Carroll, *Through the Looking Glass*. 1862

1. INTRODUCTION

Medical concepts, by their nature, are complex notions. Patient descriptions about diagnoses or procedures often invoke levels of detail and chained attributes that pose complex computer-science problems for data representation. Confounding this mechanical complexity is the sheer scale and scope of concepts that can figure into medical thought, ranging from molecular variance to sociologic environments, aptly illustrated by Blois (1988) two decades past. This breadth can be aggravated by invoking concepts and terms that can probe the depths of present knowledge, bordering unto arcane realms of science and clinical practice with limited understanding and less experience. Furthermore, a serious tension remains between making such expressions readable and understandable by humans while attempting to address the increasing need for machine-interpretable expressions that leverage computerized knowledge and decision support. Finally, had we perfect knowledge of medicine, our patients, or the spectrum of sciences medicine invokes, the task of consistently representing patient information would be hard enough. Sadly, oftentimes we struggle with incomplete information, partial understanding, and flawed models. Managing this morass to address efficiently and effectively the multiple uses of medical concepts is hard (Rector, 1999).

1.1 Use-cases

A purely abstract discussion of medical concept representation is unbounded. Enumeration of neurotransmitter molecules on a motor endplate receptor bears comparison to angels on pin-heads for clinical purposes. On the other hand, excessive reductionism may suggest a need for little more than rudimentary collections of medical terms served up as pick-lists, though this perspective typically has more to do with hiding complexity than eliminating it.

Use-case definitions serve as a practical means for defining beginning assumptions and scope as well as bounding problem spaces. Thus to frame the context of this chapter with respect to medical concept representation, some broad-based use-cases are outlined.

1.1.1 Information Capture

The most clinically familiar circumstance of representing medical concepts is documenting patient findings, conditions, interventions, and

outcomes. This documentation ranges from the unstructured dictations of progress notes and summaries to the thoughtful management of fully encoded problem lists, flow sheets, and encounter codes. Care providers thus assimilate information about patients, make inferences from patterns of observation, and re-express observations and conclusions. These expressions comprise a kind of concept representation, though not always formalized.

Medical concept representation typically implies a formal or at least machinable manifestation of clinical information. However, for practical purposes the field covers the full spectrum of information capture, including natural language. The problem of mapping natural language expressions to controlled terminologies is a topic unto itself. However, many of the challenges outlined in this chapter pertain equally to natural language expressions and more formal manifestations.

1.1.2 Communication

Information about patients, specimens, and experiments often needs to be transferred among providers or within a health care enterprise. Transmission media range from the non-machinable "fax" of text images to a highly structured clinical message conforming to the HL7* Version 3 information model. Typically, medical concepts inherit the characteristics associated with their capture, with respect to detail, formalization, and structure. However, standard message protocols such as those developed by HL7 may impose degrees of formalization that require transformation of medical concepts into highly structured representations.

Communication of medical concepts occurs for a purpose and requires that the recipient can use the information. When the recipient is a human being, seeking to read an historical medical record, these concepts may require minimal information. However, for the machine-processable transfer of electronic medical records between a referring physician and a tertiary medical center, as envisioned by the National Health Information Infrastructure (Yasnoff et al., 2004), a higher degree of interoperability is needed. Similarly, for the systematic processing of drug orders to avoid medical errors within an enterprise, a machine-interpretable representation of concurrent medications and medical problems must be achieved. This latter communication begins to overlap the formalization requirements for decision support, as described here.

* http://www.hl7.org

1.1.3 Knowledge Organization

The organization of medical knowledge is among the oldest applications of classification, dating to Aristotle's efforts in biology and formal descriptions (Pellegrin, 1986). The subsequent history (Chute, 1998; Chute, 2000) of medical concepts tells the story of increasingly detailed classifications from the haphazard collection of causes of death in the 16th century London Bills of Mortality (Graunt, 1939) to the emergence of large description logic-based terminologies (Baader et al., 2002) such as SNOMED CT†.

The explosion of modern biomedical ontologies (Smith and Rosse, 2004) provides what Alan Rector has described as a "conceptual coat-rack" for medical knowledge that knowledge authors and users find irresistible. Furthermore, the boundaries between representing concepts in an ontology using acyclic graphs and complex relations begin to blur the distinction between knowledge representation and concept organization. This realization was articulated nearly 40 years ago (Lindberg et al., 1968; Bloise, Tuttle, and Sherertz, 1981).

1.1.4 Information Retrieval

Most information sources have an indexing infrastructure that facilitates rapid and accurate retrieval. The oldest biomedical database that supports indexed retrieval is Medline/PubMed, for which the MeSH (Medical Subject Heading) vocabulary (Nelson et al., 2004) was created and is maintained. Virtually every user of the medical literature has encountered MeSH concepts, if only indirectly. Most user interfaces to literature retrieval tools translate natural text entries into MeSH concepts and then retrieve medical journal articles that have these MeSH codes or their hierarchical children (concept explosion or recursive subsumption).

For clinical data, classifications such as ICD-9-CM‡ serve an indexing role roughly corresponding to MeSH. However, ICD codes in most countries are applied for billing purposes and may not accurately reflect the underlying clinical content (Chute et al., 1996).

Whether to use natural language or coded data is an old question (Cote, 1983), though most modern practitioners recognize that any subsequent inferencing on retrieved information, using statistical regressions or machine learning techniques, must ultimately categorize or "bin" the data. Taken to the limit, such categorization defines concept classification systems. The

† http://www.snomed.org
‡ http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm

topic of medical information retrieval is addressed more completely by Hersh in this volume and elsewhere (Hersh, 2003).

1.1.5 Decision Support

Helping clinicians make better decisions all the time is arguably the ultimate goal of computer-assisted decision support systems. However, in order for such systems to work, the knowledge resources that drive decision rules must share the terms and concepts used by clinicians to describe the patient. For example, decision rules made to operate on sulfa drugs may not "fire" if they do not recognize drug trade names (e.g., Bactrim®) as equivalent. Failure to recognize semantic equivalence is a more serious challenge when confronting the myriad expressions and terms that can describe a disease. This equivalence can be daunting when a concept is fully represented using terminology composition in one setting but constitutes a combination of terms in specific fields where the information model or field semantics modify meanings in another setting. The classic example of this circumstance is "family history of heart disease" vs. "heart disease" in a field labeled "family history."

The Arden Syntax[§], a popular standard for expressing medical logic modules and decision support rules, suffers from an incomplete specification of rule triggers and vocabulary semantics. Often called the "curly braces problem" (Choi, Lussier, and Mendoca, 2003) after the typographical brackets used to contain trigger concepts and rule-logic terms, implementers of a decision rule published in the Arden Syntax were left to their own devices to interpret exactly what events and codes in their own organization best correspond to the concepts within the curly braces. This semantic challenge highlights the importance of shared concept representation among logic-rule authors, implementers, and users.

2. CONTEXT

The settings of use often define as much about concepts as any surface form or text string might convey. The famous linguistic example of contextual syntax is "Time flies like an arrow, but fruit flies like a banana." This example illustrates the profound changes of meaning that context can have on words, terms, and expressions. The biomedical domain, while often more structured than general language, does not escape the influence of context on the representation or interpretation of concepts.

[§] http://www.hl7.org/Special/committees/Arden/arden.htm

The definitively cited work on context, language, and concepts remains Ogden and Richards' 1923 opus, *The Meaning of Meaning* (Ogden et al., 1923). They describe the classic "semantic triangle" which distinguishes a purely abstract thought or human concept, a referent object in the real world, and language symbols we might culturally share to refer to this concept. Invoking the Shakespearian metaphor of a "'rose' by any other name…," Ogden outlines that the shared cultural context of a rose – merely a pretty flower or a symbol of love – dictates its interpretation. These shared cultural assumptions are little different in health care, though comprehensive medical concept representations in the guise of a fully-specified HL7 message leaves less context to assumption and more to explicit assertion.

2.1 Concept Characteristics

Disease descriptions exist along many axes of characteristics, defining continua of expression. These axes have implications for managing concept representation and interpreting concept instances.

- Certainty – Clinicians document medical concepts throughout a care episode, including periods when they are unsure of their own speculations. Clinical assertions range from differential diagnoses, which include broad possibilities, probable but uncertain observations, to final diagnoses (though these too are often revised). Hence, many concepts in patient records may comprise more noise than fact.
- Etiologic Precision – Diagnostic statements are fraught with vagueness, syndromic generalization, and final common pathway manifestations attributable to multiple causes. Consider the label "congestive heart failure," which exhibits myriad etiologies though shared clinical outcomes. Many medical concepts exhibit substantial clinical heterogeneity. Contrasting such vagueness is the emergence of an increasing number of clinical characterizations that correspond to precise molecular variations (Scriver, 2001; McKusick), such as hemoglobinopathies or specific tumors. Indeed, the entire genomic revolution will inevitably transform our understanding of disease and etiology in a manner analogous to the effect of the germ theory of disease.
- Granularity (specificity) – Disease hierarchies are not just the province of classifications, but find expression in clinical descriptions. There is a profound difference between a problem list entry of "cancer" and one that specifies "Stage IIb squamous cell carcinoma of the right upper lobe with metastatic extension to the liver…" Reference to a "granular" description implies a detailed expression, often as a composition. More

specific terms can be distinguished without composition (e.g., the granular “aortic insufficiency” contrasting with “heart disease”).

- Completeness – specific use-cases often determine how completely clinical descriptions are expressed. Routine outpatient office visits may exhibit a limited amount of disease detail when compared with the detail provided through an elaborate clinical trial protocol. The boundary between completeness and granularity is often determined by how the information is represented between vocabulary expressions vs. information model structures.

2.2 Domains

The professional language or jargon of medicine differs markedly from general English (or any other natural human language). While health professionals doubtless share substantial biomedical sub-language elements, there is important sub-specialization by medical specialty. Neonatologists do not fully share the language of psychiatrists; similar contrasts could be drawn between the language of cardiologists and pathologists, radiologists and clinical pharmacologists, rehabilitation specialists and oncologists, and so on. These distinctions define concept domains, although domains are by no means limited to differences in clinical sub-specialties.

A palpable way to illustrate these distinctions is to examine how certain specialties might disambiguate simple and common abbreviations. The following table expands differently the abbreviation MS by some domain specialties. What is remarkable is that these expansions occur consistently *within* domains, but almost always inconsistently *among* domains. The exercise is equally repeatable with MI, MR, and countless other abbreviations.

Table 6-1. Domain-specific expansion of "MS"

Domain specialty	“MS” abbreviation expansion
Cardiology	mitral stenosis
Neurology	multiple sclerosis
Anesthesia	morphine sulfate
Obstetrics	magnesium sulfate
Research science	manuscript
Physics	millisecond
Education	Master of Science
U.S. Postal Service	Mississippi
Computer science	Microsoft
Correspondence	female name prefix

Domain-specific term disambiguation is not restricted to abbreviations. The NLM's UMLS[**] contains six meanings for "cold." One is an abbreviation expansion (chronic obstructive lung disease). However, each of these meanings carries a unique concept identifier (CUI) within the UMLS that can be invoked to represent a context-independent statement. Furthermore, concepts can be fully expressed in language to avoid ambiguity, although most human interfaces find fully disambiguated text expansions tedious at best and sometimes insufferable. Using widely understood shorthand expressions *within* a domain for human consumption is a practice not likely to languish anytime soon.

2.3 Structure

The meaning of a term is as much influenced by the company it keeps (structural context) as by who uses it (domain). However, the expression of structural context has a dual nature in medical concept representation, as illustrated in the figure below. Specifically, highly detailed, granular vocabulary expressions can be composed which express a complex notion illustrated by the vocabulary composition view. Semantically identical assertions can be expressed using shorter vocabulary elements within a specific information model that conveys the additional semantics – in this case, the qualification of "family history."

When one begins to deal with more complex information models and more expressive vocabulary spaces, the problem worsens. The following table is adapted from material suggested by David Markwell of the UK at the inaugural TermInfo meeting held at National Aeronautics and Space Administration (NASA) in Houston, TX during August 2004. This series of meetings was convened to examine the spectrum of concept modeling that can exist between terminology models (such as SNOMED or GO) and information models (such as HL7 reference information model (RIM) or caBIO), and in particular where these models generate a semantic overlap. The table highlights alternative ways of modeling the same information by using HL7 RIM and the SNOMED CT context model.

Table 6-2. HL7 RIM and SNOMED CT Context Model

HL7 RIM	SNOMED CT Attribute
targetSiteCode(Observation)	"finding site"
targetSiteCode(Procedure)	"procedure site"
methodCode(Observation & Procedure)	"method"
approachSiteCode(Procedure)	"approach," "access"
priorityCode(Act)	"priority"

[**] http://www.nlm.nih.gov/research/umls/

The conclusion, almost inescapably, is that there is no one correct way to represent complex medical concepts. Invoking higher-level information models such as the HL7 RIM or even just a "family history" box has equivalent validity and semantics to composition expressions built using vocabulary models and syntax. If both are valid, then what is the problem?

The resolution of complex, semantically equivalent expressions that differ in their allocation of meaning to an information model or compositional vocabulary expressions is difficult. Establishing semantic equivalence between such hybrid representations – or even their purely modeled archetypes of complete information model or vocabulary expression – is an under-developed research problem. Few solutions exist, and none scale to the scope of problems encountered in real-world clinical expressions. The practical implication is that virtually all use-cases that require communication or consistent recognition of content by a recipient (as in decision support) will fail, should care not be taken to negotiate the allocation of semantics between information and vocabulary models.

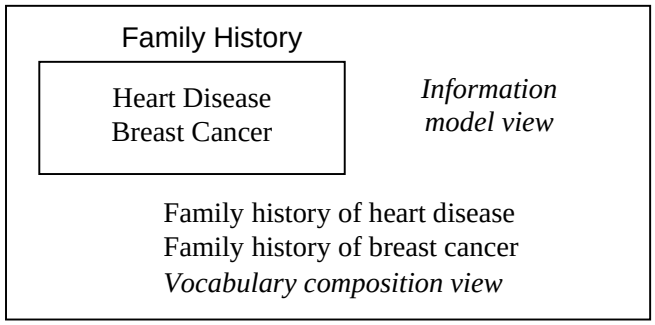

Figure 6-1. Information model view and vocabulary composition view.

3. BIOMEDICAL CONCEPT COLLECTIONS

3.1 Ontologies

Philosophers recoil at the pluralization of ontology. Originally, the term ontology referred to the consideration of what kinds of entities comprise reality. Computer scientists, in the era of artificial intelligence and knowledge representation, co-opted the term to mean an organization of concepts in domains, which might encompass medical concepts or enumerations and relations among Boeing 777 aircraft parts. Gradually,

criteria discriminating formal ontologies from ordinary hierarchies of concepts included the requirement that ontologies exhibit internal consistency, acyclic polyhierarchies, and computable semantics. Within medicine, the pioneering work of Rector and the Galen project (Rector, Nowland, and the Galen Consortium, 1993) illustrated how medical concepts could be represented as a formal ontology and demonstrated applications where this formalized representation mattered (Rector, Rossi Mori, and Consorti, 1993).

Modern biomedical ontologies are becoming synonymous with concept collections assembled using description logics (Baader and Nutt, 2002). The venerable SNOMED has evolved to incorporate description logics which include role restrictions (Spackman et al., 2002). Concept collections pertaining to basic biology, such as the Gene Ontology, while criticized for lacking many formalizations now expected of terminology bearing the ontology banner (Smith, Williams, and Schulze-Kremer, 2003), are also evolving to become a semantically computable resource (Wroe et al., 2003).

The novel promise of ontologies is their ultimate connection as a distributed system of interlocking conceptual schema. For example, when a LOINC term invokes a drug within a drug-sensitivity evaluation, semantic interoperability is enhanced if there emerges agreement that the NDF-RT (Chute et al., 2003) (National Drug Formulary – Reference Terminology; a UMLS source vocabulary) would form the basis for drug references. Similarly, the designation of a common anatomy terminology that could span the spectrum of use-cases across biology and medicine would greatly enhance our ability to consistently create and interpret biomedical concepts; a leading candidate for a common anatomical ontology appears to be the Foundational Model of Anatomy developed by Smith and Rosse (2004).

The development and deployment of ontologies is being greatly accelerated by the emergence and adoption of the Protégé ontology editor developed by Musen and colleagues (Noy et al., 2003). Closely coupled with this tooling is agreement upon an expression syntax for description logics, published by the WC3. Called OWL (the Ontology Web Language), this standard (McGuinness and Harmelen, 2004) was approved only recently but has already penetrated the ontology authoring community almost completely. As of this writing, the OWL extensions to Protégé (Knublauch, 2004) provide the best authoring and editing environment for ontology development available anywhere; that this NIH-funded effort is now available[††] with an open-source license should further accelerate the quality and number of well-formed ontologies in biomedicine.

[††] http://protege.stanford.edu; funded in part by NIH grant P41-LM07885 to Mark Musen, MD PhD

3.2 Vocabularies and Terminologies

Cimino provides a detailed description of vocabularies and terminologies in this volume. For the purposes of this discussion, it is useful to distinguish vocabularies and terminologies from ontologies. Simplistically, vocabularies and terminologies are less formal than ontologies, uniformly lacking logical descriptions that serve to computationally define terms. As a practical matter, most large ontologies contain a sizable fraction, if not a majority, of "primitive" terms undefined by description logic formalisms – terminologies remain the major mode for biomedical terminologies, if not at some levels the exclusive mode for the present.

There is no commonly accepted distinction between a vocabulary and terminology, though many adherents in the field might suggest that terminologies have associated codes and hierarchies while a simple vocabulary may comprise little more than a bag of words. However, invocation of the moniker "controlled vocabulary," which may imply more formality than exists in a terminology, renders this tenuous differential inconsistent.

Vocabularies and terminologies are often described by intended role, though few adhere to these role expectations. The most common distinctions among terminology uses are:

- Entry Terminologies: specifically constructed to provide familiar and common terms and phrases readily recognized by humans. These term collections often sacrifice precision and rigor in favor of familiarity and jargon.
- Reference Terminologies: semi-formal representations of terms and concepts intended for machine interpretation.
- Administrative Terminologies: higher-level classifications which aggregate clinical findings for particular administrative purposes.

Common vocabularies and terminologies include LOINC, CPT (Current Procedural Terminology), HL7 Vocabulary Tables (over 100 of them), and NDC drug codes. People familiar with these systems can recognize wide variations in structure, quality, and consistency of these concept libraries. What makes a terminology good is beyond the scope of this work, though systematic evaluations of common terminologies exist (Chute et al., 1996; Campbell et al., 1997), as do generalized discussions of what makes a good terminology (Chute, Cohn, and Campbell, 1998; Cimino, 1998). The reality remains that most terminologies fail to adhere to good design principles, suffering from the recycling of abstract codes to inconsistent hierarchies to ambiguous groupings of concepts.

3.3 Aggregation and Classification

Historically, significant tension existed between the terminology and classification communities (Cote, 1983; Ingenerf and Giere, 1998), with each maintaining the advantages of their use-case. However, recent thinking has established their mutual advantages along a continuum of granularity or specificity (Chute, Cohn, and Campbell, 1998). Medical information, by its nature, is highly detailed. Hence a need for concept systems or terminologies that can reasonably capture highly detailed information will always exist. On the other hand, many use-cases work best with highly grouped data. Examples include public-health statistics, reimbursement categories, or administrative groupings of patients.

High-level aggregation systems, such as ICD-9-CM, have been unjustly criticized for not having enough granularity to function in decision support or clinical retrieval use-cases. The complaint is accurate but the criticism unjustified because high-level classifications such as ICD-9-CM were *never intended* to function as detailed terminologies. If criticism is to be made, it should be of Electronic Health Record (EHR) vendors and most implementing providers who insist on using ICD-9-CM for use-cases such as patient problem lists and clinical decision support triggers that more properly demand detailed terminologies, such as SNOMED-CT.

However, the specter of double-coding clinical findings, diagnoses, procedures, or adverse events, once in a detailed terminology and again in a required or mandated classification, reasonably discourages best-coding practice. Few providers have the resources to appropriately code cases for reimbursement and quality oversight, never mind code them again for clinical applications. Early drafts of the PITAC (President's Information Technology Advisory Committee) Report on Health Information Technology went so far as to suggest that providers code just once, in a detailed terminology, and that secondary re-use of clinical data be facilitated by appropriate mapping to requisite classifications, such as the newly drafted ICD-10-CM. The final version (U.S. President's Information Technology Advisory Committee, 2004) of this report, however, provided a much more balanced perspective on the important roles that high-level classifications can play, coupled with the many practical difficulties of accurately mapping detailed clinical terms to complex classifications.

Kent Spackman, editor of SNOMED-CT, proposed (personal communication) that mapping from detailed terminologies to complex classification would provide more reliable and consistent coding. However, he points out that to be done correctly, the coding rules of a classification, such as ICD-9-CM, must be made explicit and machineable. Most classifications rely on indentations, typographic conventions, index entries,

and established professional coding lore as a basis for conveying the rules of coding. These rules can be quite elaborate, specifying complex inclusion and exclusion criteria for assignment to a specific code. As a simple example, pre-eclampsia is distinguished from ordinary hypertension in ICD-9-CM by obviously requiring female gender, pregnancy, and renal involvement.

These coding rules could define "Aggregation Logics," and should be published as machine-readable logic rules by developers of classifications. The analogy is often made to "Grouper" rules, by which collections of ICD codes are grouped into higher-level DRGs (Diagnostic Related Groups) by computer algorithms. Aggregation Logics would fill the gap between detailed clinical expressions and the intermediate classifications, such as ICD-9-CM or ICD-10-CM, when it becomes adopted. The point is to avoid duplicate coding by providers, and consistent with Spackman's assertion, likely provide more reliable and consistent coding into ICD-level classifications.

3.4 Thesauri and Mappings

3.4.1 The UMLS

No discussion of biomedical terminology and concept representation would be complete without mention of the Unified Medical Language System (*op. cit.*). Originally intended to serve as a Rosetta Stone to suggest translations among terminologies (Humphreys and Lindberg, 1989), it has taken a more practical role as the major semantic thesaurus of biomedical terms. The UMLS is comprised of over 100 separate terminology sources, including SNOMED CT, MeSH, and ICD-9-CM. However, it does not contain formal description logic assertions across terms from different vocabularies, though hierarchical assertions, broader/narrower relations, and "other" relationships are meticulously mapped and curated by human editors for the entire corpus.

The 2004 release of the UMLS Metathesaurus saw the most dramatic change in the file structures and formats of the UMLS since its original release in 1988. To accommodate the complex description logic assertions of SNOMED CT, the NLM introduced a Rich Release Format (RRF) (National Library of Medicine, 2003), which for the first time promised "source transparency." The intention was to permit users of the UMLS to extract terminologies from the Metathesaurus in a format that would transparently reflect the original content of a particular terminology. Previously, the UMLS formatting process resulted in a "lossy" information transfer. The modern vision of the UMLS, to become the definitive source

and publication format for major biomedical terminologies, is thereby greatly advanced.

3.4.2 Word-Level Synonymy

An emerging requirement for natural language thesauri is not presently served by the UMLS, though it is approximated for general English by resources such as WordNet (Fellbaum, 1998). Consider the retrieval use-case for Renal Cancer when data may have been recorded as Kidney Cancer. The UMLS happens to include explicit English synonyms that map these two phrases, but this is not the case for all word-level synonyms and permutations that one might imagine in biomedical concepts.

The public sharing of word-level concept clusters has been widely proposed (Solbrig et al., 2000), and indeed some generalized methods for creating and evolving them have been explored (Pakhomov, Buntrock, and Chute, 2004). The broad creation, shared maintenance, and coordinated use of consensus-driven thesauri of common synonyms will be a great advance toward linking phrases entered by providers with elements of controlled terminologies. These resources, in a second generation of curation, could also include degrees of pleisionymy. Ultimately, these thesauri can be married with ontologies and terminologies to provide a horizontal (synonym) and vertical (terminologies) component to medical concept representation and retrieval (Chute, 2002).

4. STANDARDS AND SEMANTIC INTEROPERABILITY

Medical concepts, once expressed, must be understood by people or machines. The context of concept assertion can overlay additional semantics that must be understood. Fully specified messaging environments, such as HL7 or caBIO[‡‡], can carry sufficient information to explain context, but there is no replacement for agreed-upon content standards, to wit common vocabularies.

In the United States, the Federal eGov initiatives have spawned the Consolidated Health Informatics (CHI) set of standards[§§]. Working in concert with the NCVHS, the CHI working groups have proposed terminology and interchange standards that would be required for use among US Federal agencies. Intended to define a critical-mass tipping point for the

[‡‡] http://ncicb.nci.nih.gov/core/caBIO
[§§] http://www.whitehouse.gov/omb/egov/gtob/health_informatics.htm

general US health care economy, the explicit intention is that such Federal leadership would define de facto a common basis for content standards. While the CHI proposals are still new as of this writing, the intended effect appears to be taking place. Evidence for this can be seen in the NLM contract to HL7 (HHS N276 2004 43505C) to ensure that all HL7 vocabulary tables are CHI-compliant.

One may conclude that substantial progress and tangible resources have emerged in the past few years to support the consistent and comparable representation of medical concepts for a broad spectrum of use-cases. The rapid adoption of ontology languages such as OWL, their subsequent availability in high fidelity within the UMLS, and the active negotiation and specification of what contextual information belongs in an information model vs. a terminology model bring increasing problems of robust solutions. The common use of highly detailed and semantically coherent medical messages and retrievals is not yet realized, but progress has been dramatic in the past five years. The clichéd refrain that more work needs to be done certainly pertains, but that work is now more palpably satisfying and is vectoring toward consensus solutions and practical standards specifications.

5. ACKNOWLEDGEMENTS

I am grateful to Harold Solbrig, Mark Musen, and Alan Rector for fruitful discussions that contributed to my understanding of these issues. This work is funded in part by R01 LM007319.

REFERENCES

Baader, F., Calvanese, D., McGuinness, D. L. et al. (Eds.). (2002). *The Description Logic Handbook: Theory, Implementation and Applications*, Cambridge University Press.

Baader, F. D. and Nutt, F. W. (2002). "Basic Description Logics," in F. Baader et al. (Eds.), *The Description Logic Handbook*, Cambridge University Press, 47-100.

Blois, M., Tuttle, M., and Sherertz, D. (1981). "RECONSIDER: A Program for Generating Differential Diagnoses," in *Proceedings of the 5th Annual Symposium on Computer Applications in Medical Care*, Washington, D.C., Nov. 1-4, 1981, 263-268.

Blois, M. S. (1998). "Medicine and the Nature of Vertical Reasoning," *New England Journal of Medicine*, 318(13), 847-851.

Campbell, J. R., Carpenter, P. C., Sneiderman, C. et al. (1997). "Phase II Evaluation of Clinic Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity," *Journal of the American Medical Informatics Association*, 4(3), 238-251.

Choi, J., Lussier, Y. A., and Mendoca, E. A. (2003). "Adapting Current Arden Syntax Knowledge for an Object Oriented Event Monitor," in *AMIA Annual Symposium Proceedings*, 814.

Chute, C. G. (2002). "The Horizontal and Vertical Nature of Patient Phenotype Retrieval: New Directions for Clinical Text Processing," in *Proceedings of the AMIA Annual Fall Symposium*, 165-169.

Chute, C. G. (2000). "Clinical Classification and Terminology: Some History and Current Observations," *Journal of the American Medical Informatics Association*, 7(3), 298-303.

Chute, C. G. (1998). "The Copernican Era of Healthcare Terminology: A Re-centering of Health Information Systems," in *Proceedings of the AMIA Symposium*, 68-73.

Chute, C. G., Carter, J. S., Tuttle, M. S. et al. (2003). "Integrating Pharmacokinetics Knowledge into a Drug Ontology: As an Extension to Support Pharmacogenomics," in *Proceedings of the AMIA Symposium*, 170-174.

Chute, C. G., Cohn, S. P., and Campbell, J. R. (1998). "A Framework for Comprehensive Health Terminology Systems in the United States: Development Guidelines, Criteria for Selection, and Public Policy Implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures," *Journal of the American Medical Informatics Association*, 5(6), 503-10.

Chute, C. G., Cohn, S. P., Campbell, K. E. et al. (1996). "The Content Coverage of Clinical Classifications," *Journal of the American Medical Informatics Association*, 3(3), 224-233.

Cimino, J. J. (1998). "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century," *Methods of Information in Medicine*, 37(4-5).

Cote, R. (1983). "Editorial: Ending the Classification Versus Nomenclature Controversy," *Medical Informatics*, 8(1), 1-4.

Fellbaum, C. (1998). "WordNet: An Electronic Lexical Database," in *Language, Speech, and Communication*, Cambridge, Mass: MIT Press.

Graunt, J. (1939). *Natural and Political Observations Made Upon the Bills of Mortality; London, 1662*, Baltimore, MD: The Johns Hopkins Press.

Hersh, W.R. (2003). *Information Retrieval: A Health and Biomedical Perspective*, New York: Springer.

Humphreys, B. L. and Lindberg, D. A. B. (1989). "Building the Unified Medical Language System," in *Symposium on Computer Applications in Medical Care*, 13, 475-480.

Ingenerf, J. and Giere, W. (1998). "Concept-oriented Standardization and Statistics-oriented Classification: Continuing the Classification Versus Nomenclature Controversy," *Methods of Information in Medicine*, 37(4-5).

Knublauch, H. (2004). "The Protégé OWL Plugin," in *7th International Protégé Conference*, Bethesda, MD, http://protege.stanford.edu/conference/2004/index.html

Lindberg, D. A. B., Rowland, L. R., Buck, C. R. et al. (1968). "CONSIDER: A Computer Program for Medical Instruction," in *Proceedings of the 9th IBM Med. Symposium*, White Plains, New York: IBM.

McGuinness, D. L. and Harmelen, Fv. (2004). "*OWL Web Ontology Language: Overview*," W3C, http://www.w3.org/TR/owl-features/

McKusick, V. A. *OMIM - Online Mendelian Inheritance in Man*, Bethesda: National Center for Biotechnology Information, NIH/NLM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

National Library of Medicine. (2003). *MLS Metathesaurus Rich Release (MR+) Format*, Bethesda, MD: National Institutes of Health, http://www.nlm.nih.gov/research/umls/white_paper.html

Nelson, S. J., Schopen, M., Savage, A. G. et al. (2004). "The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation," in *Medinfo*, 67-9.

Noy, N. F., Crubezy, M., Fergerson, R. W. et al. (2003). "Protege-2000: An Open-source Ontology-Development and Knowledge-Acquisition Environment," in *Proceedings of the Annual AMIA Symposium*, 953.

Ogden, C. K., Richards, I. A., Malinowski, B. et al. (Eds.) (1923). *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*, London:Routledge & Kegan Paul.

Pakhomov, S. V., Buntrock, J. D., and Chute, C. G. (2004). "Using Compound Codes for Automatic Classification of Clinical Diagnoses," in *Medinfo*, 411-415.

Pellegrin, P. (1986). *Aristotle's Classification of Animals: Biology and the Conceptual Unity of the Aristotelian Corpus*, Berkeley: University of California Press.

Rector, A. L. (1999). "Clinical Terminology: Why is it So Hard?" *Methods of Information in Medicine*, 38(4-5), 239-252.

Rector, A. L., Nowland, W., and The Galen Consortium. (1993). "The GALEN Project," *Computer Methods and Programs in Biomedicine*, (45), 75-78.

Rector, A. L., Rossi Mori, A., Consorti, M. F. et al. (1998). "Practical Development of Re-usable Terminologies: GALEN-IN-USE and the GALEN Organization," *International Journal of Medical Informatics*, 48(1-3), 71-84.

Scriver, C. R. (2001). *The Metabolic and Molecular Bases of Inherited Disease*, 8th ed. New York: McGraw-Hill. 4 vols, (xlvii, 6338, I-140 p.).

Smith, B. and Rosse, C. (2004). "The Role of Foundational Relations in the Alignment of Biomedical Ontologies," in *Medinfo*, 444-448.

Smith, B., Williams, J., and Schulze-Kremer, S. (2003). "The Ontology of the Gene Ontology," in *AMIA Annual Symposium Proceedings*, 609-613.

Solbrig, H., Elkin, P., Ogren, P. et al. (2000). "A Formal Approach to Integrating Synonyms with a Reference Terminology," in *Journal of the American Medical Informatics Association Symposium Supplement*.

Spackman, K. A., Dionne, R., Mays, E. et al. (2002). "Role Grouping as an Extension to the Description Logic of Ontylog, Motivated by Concept Modeling in SNOMED," in *Proceedings of the AMIA Symposium*, 712-6.

United States. President's Information Technology Advisory Committee, and the National Coordination Office for Information Technology Research and Development. (2004). "Revolutionizing Health Care through Information Technology Report to the President," Arlington, VA: National Coordination Office for Information Technology Research and Development, http://www.itrd.gov/pitac/reports/20040721_hit_report.pdf

Wroe, C.J., Stevens, R., Goble, C. A. et al. (2003). "A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL," in *Pacific Symposium on Biocomputing*, 624-635.

Yasnoff, W. A., Humphreys, B.L., Overhage, J. M. et al. (2004). "A Consensus Action Agenda for Achieving the National Health Information Infrastructure," *Journal of the American Medical Informatics Association*, 11(4), 332-338.

SUGGESTED READINGS

Baader, F., Calvanese, D., McGuinness, D. L. et al. (Eds.) (2002) *The Description Logic Handbook: Theory, Implementation and Applications*, Cambridge University Press.

The definitive textbook which outlines the history and current state-of-the-art for Description Logics. Since Description Logics form the basis of modern ontologies, familiarity with this technology is increasingly required for mastery of concept representation.

Rector, A. L. (1999). "Clinical Terminology: Why is it so Hard?" *Methods of Information in Medicine*, 38(4-5), 239-252.
An outstandingly concise and complete exposition on the terminology problem in health care, effectively refuting commonly held expectations that health terminology should be trivial.

Chute, C. G. (2000). "Clinical Classification and Terminology: Some History and Current Observations," *Journal of the American Medical Informatics Association*, 7(3), 298-303.
A brief history of medical classification and description, providing background and context for the evolution of thinking and practice in health classifications through the last century.

ONLINE RESOURCES

http://umlsks.nlm.nih.gov/
The home side of the NLM's Unified Medical Language Systems

http://informatics.mayo.edu
The specification and open-source for the LexGrid project, terminology editor, and Common Terminology Services (from HL7).

http://protege.stanford.edu
The most widely used ontology editor, Protégé, and related resources.

http://www.co-ode.org/
The Collaborative Open Ontology Development Environment home page, including tutorials and resources

QUESTIONS FOR DISCUSSION

1. What are the relative roles of terminology models and information models in representing complex medical expressions?

2. What is the distinction between representing information and aggregating information? Specifically, what are the relative roles and relationships among terminologies and classifications?

3. How might a spectrum of secondary data uses, such as decision support, quality improvement, biomedical research, or administrative aggregation, impact information representation and display?

4. How might the retrieval of information be affected by differing ways of representing it? Specifically include discussion of granularity, detail, aggregations (lumping), or context?