High-quality, Standard, Controlled Healthcare Terminologies Come of Age

J. J. Cimino

Laboratory for Informatics Development, The Clinical Center of the National Institutes of Health, Bethesda, Maryland, USA

Controlled terminologies play an integral role in biomedical informatics applications whenever such applications use information to carry out symbolic reasoning. Early, prominent examples include Mycin [1], which included terms for clinical findings in its rules and presented diagnosis and treatment terms in its recommendations, and the Internist-1 system [2], which matched a list of symptom and physical finding terms against a knowledge base of disease terms. While PubMed [3], at the US National Library of Medicine (NLM), can carry out literature retrieval with simple text searching, it also makes use of controlled terms to focus that retrieval. Today, everything from automated health surveys to electronic health records makes use of some kind of controlled terminology to condense users' input into a set of symbols that can be recognized and manipulated.

Standardized controlled terminologies are much less ubiquitous in health care applications. The Medical Subject Headings (MeSH) used in PubMed is one such example, although it is primarily a standard within the NLM. The most ubiquitous standard terminologies in health care are the World Health Organization's International Classification of Diseases (ICD) and its variations, including the 9th edition (ICD-9), the 10th edition (ICD-10) and assorted Clinical Modifications thereof (e.g., ICD-9-CM, ICD-9-AM, ICD-9-CA, and ICD-10-CM). The use of ICDs can be traced back over 100 years, with increased international use after the World Health Organization began requiring United Nations member countries to use them for reporting mortality statistics. Further incentives for use of the ICD terminologies derived from requirements of healthcare payors (insurance companies and government agencies) to include ICD codes (particularly ICD-9-CM) when billing for patient care. These requirements were often a major impetus for hospitals to purchase computers and record patient data electronically [4].

It is not surprising, then, that patient data recorded with ICD-9-CM and its relatives have often been reused for a variety of purposes, including administrative functions, epidemiologic studies, research subject recruitment, interventional protocols, and clinical decision support systems. After all, the data are already being collected and they are ubiquitous, making them low-hanging fruit for those who seek data in a standard, controlled form but have few resources to generate them. There are, however, some drawbacks to using ICD-coded data for purposes other than their original intent. Some of the problems are related to the process by which the data are captured, with error rates as high as 90% in some studies (see for example, a review by O'Malley and colleagues [5]). However, some of the problems with such data relate to the design aspects of the terminologies used to record them. I will use a few examples from ICD-9-CM to illustrate the kinds of limitations that can be traced to terminology design.

Any controlled terminology will necessarily lack the richness of detail available from the vocabulary of a natural language; this loss of this detail is one of the trade-offs for having data in a computable form. However, the level of detail permitted with

Correspondence to:
James J. Cimino, MD
Chief, Laboratory for Informatics Development
The Clinical Center of the National Institutes of Health
10 Center Drive
Bethesda, MD 20892
USA
E-mail: ciminoj@cc.nih.gov

Table 1 Hypothetical shock mortality data showing ICD-9-CM coding in the 2001–2003 time period. Note that although data for Septic Shock and Hypovolmeic Shock are available separately, they are aggregated into the same ICD-9-CM code (785.59) in 2001 and 2002 (separate row at bottom); ICD-9-CM did not contain the code 785.52 in those years.

	2001				2002				2003			
	Cases	Deaths	Mortality	ICD-9-CM	Cases	Deaths	Mortality	ICD-9-CM	Cases	Deaths	Mortality	ICD-9-CM
Cardiogenic Shock	300	30	10.0%	785.51	350	34	9.7%	785.51	300	28	9.3%	785.51
Septic Shock	200	30	15.0%	785.59	220	30	13.6%	785.59	210	25	11.9%	785.52
Hypovolemic Shock	100	20	20.0%	785.59	120	22	18.3%	785.59	110	19	17.3%	785.59
All 785.59	300	50	16.7%		340	52	15.3%		110	19	17.3%	

ICD codes is particularly limited by the restrictive nature of the numbering system used to create term identifiers (codes), allowing few hierarchical levels (typically, two to three) with only a small number terms at each level (typically ten). As a result, many otherwise interesting terms are forced into "not elsewhere classified" bins, simply because the coding system ran out of room. For example, the toxic effects of many individual metals are represented with metal-specific codes that begin with "985"; however, brass fumes, copper salts, and iron and nickel compounds are all included under a single code, 985.9 (*Toxic Effect of Other Specified Metals*). Perhaps this aggregation is intentional on the part of the terminology authors, but even if they had wanted to have separate codes for each metal, they could not: there are simply not enough codes available. While this loss of detail may not interfere with the original intended use of ICD-9-CM-coded data, the use of a decimal coding system results in unnecessary limits on data reuse.

Another structural limitation of the ICD terminologies is related to their organization as strict hierarchical structures in which terms can only have one lineage. Thus, infectious disease of the lung cannot fall under both *Diseases of the Respiratory System* (codes 460–519) or *Infectious and Parasitic Diseases* (codes 001–139). Bacterial pneumonias are found in the former section, except pulmonary tuberculosis, which is found in the latter section. This can present problems for users of ICD-encoded data who seek to determine whether patients fall into categories of interest. For

example, a researcher trying to retrieve data on patients with tuberculosis might logically expect that such data will be one of the many codes that fall in the range of 010.00 (*Primary Tuberculosis Infection, Unspecified*) to 018.96 (*Miliary Tuberculosis, Unspecified, Tubercle Bacilli not Found by Bacteriological or Histological Examination, but Tuberculosis Confirmed by Other Methods*). A query for codes in this range is technically simple to accomplish, but would fail to identify patients with codes such as 137.0 (*Late Effects of Respiratory or Unspecified Tuberculosis*) and 647.34 (*Tuberculosis, Postpartum Condition or Complication*). Again, the problem is due to the underlying design of the terminology: a hierarchical coding system for term identifiers limits terms to a single position in the hierarchy. While a strict hierarchical arrangement is helpful for purposes that require subsetting data into mutually exclusive bins (specifically, the bins that reflect the ICD view of the world), it is a severe limitation for reuse in such functions as clinical alerting systems (where, for example, a rule might state "If the patient has a previous diagnosis of tuberculosis…").

More subtle problems arise when periodic updates to the terminology result in changes to the meanings of terms. Some changes may seem obvious, such the 1995 update to ICD-9-CM in which the name for code 664.14 changed from "*Other Immediate Postpartum Hemorrhage, Postpartum*" to "*Second-Degree Perineal Laceration, Postpartum*". However, even these changes may be hard to detect when a user of historical data has only the codes for the data

and the most recent version of the terminology.

Careful attention to the evolution of term names and codes can still fail to identify some of the more subtle, but nevertheless significant semantic drift that can occur. Consider, for example, the ICD-9-CM codes that have been used to represent nontraumatic cardiovascular collapse (shock). In 2003, the code 785.52 was introduced for coding cases of septic shock. Prior to 2003, such cases, while commonly diagnosed as such, were recorded with the ICD-9-CM code 785.59 (*Other Shock without Mention of Trauma*). Consider a hypothetical collection of patient mortality data for various forms of nontraumatic shock (▶Table 1). Although those recording the data may have dutifully used specific terms (such as "cardiogenic shock", "septic shock" and "hypovolemic shock"), the conversion of the data into ICD-9-CM produces not only a loss of detail, but a *different* loss of detail depending on the year. Someone reviewing a graphical representation of such data might be alarmed to see that the mortality rate for *Septic Shock* and *Other Shock without Mention of Trauma* both increased dramatically in 2003, even though an examination of the original data show that the opposite trends occurred[a] (▶Fig. 1). This problem of unrecognized semantic

---

[a] This apparent change in data over time has been dubbed the "Will Rogers Phenomenon", after a quote in which the American social commentator and humorist wrote "When the [destitute farmers] left Oklahoma and moved to California, they raised the average intelligence level in both states" [6].

drift is more than just theoretical; researchers who make use of multi-year ICD-9-CM data are actually unaware that the meaning of their data are changing over time [7].

The types of limitations described above apply to many non-ICD terminologies as well. Attention to terminology design, as an appropriate research area in biomedical informatics, is a relatively recent development. Two efforts stand out, in terms of research publication, as major contributors to the new focus. One of these was the Unified Medical Language System, initiated in 1986 by Don Lindberg, director of the NLM [8]. This project initially brought together informatics researchers from six universities in the US to develop methods for reconciling the distinctions between different controlled terminologies (some standard, some not) in order to facilitate information exchange and retrieval between systems using these terminologies. What quickly developed, however, was a focus on the syntactic and semantic aspects of these terminologies, leading to studies of not just how to reconcile them but how to represent them formally to expose strengths and weaknesses. This single project led to an explosion in terminologic research, with 260 publications on the UMLS alone in the first ten years between 1986 and 1996, [9] and over 766 more since then that at least mention the project [10]. Perhaps equally important, those working on the UMLS, including NLM staff, NLM contractors, and independent investigators, have gone on to continue formal research in this area.

A second, more international effort has been the work of the International Medical Informatics Association Working Group 6 (IMIA-WG6) on Natural Language, Classification and Concept Representation (now called Medical Concept Representation). Under the chairmanship of the late Jean-Raoul Scherrer and subsequently by Christopher Chute, IMIA-WG6 hosted a series of meetings that brought together researchers from around the world to present and discuss their work. These meetings not only led to many publications, including several special issues of this journal [11–13], but served as a forum for investigators who had pre-
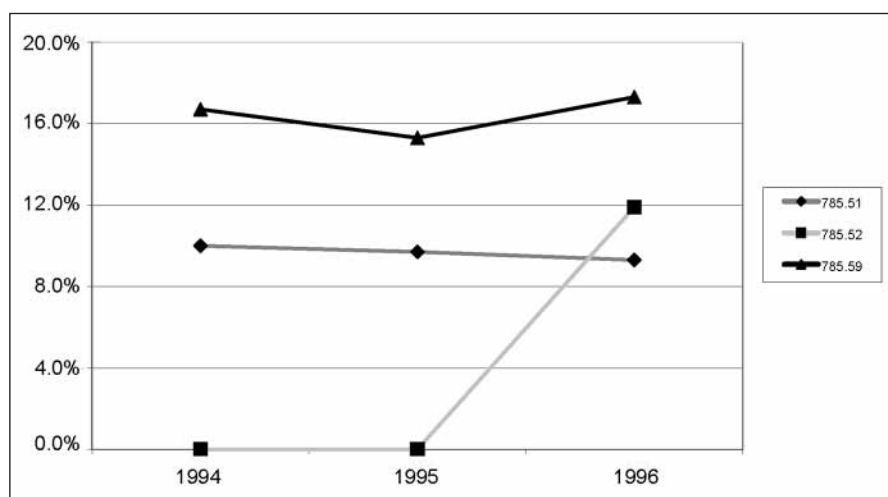

Fig. 1 Graph showing annual shock mortality, by ICD-9-CM code, based on the hypothetical data in Table 1

viously been working largely in isolation, allowing them the opportunity to establish common ground for future collaborations.

The result of these efforts (and others too numerous to review here) has been the establishment of formal principals for terminology design that have guided the development of new terminologies, such as the Logical Observations, Identifiers and Codes (LOINC)[14], but have also helped to transform well-established terminologies, such as the Systematized Nomenclature of Medicine (SNOMED)[15] and, I'm happy to report, ICD, which will adopt a formal concept-oriented model for its 11th edition (ICD-11) [16].

Even as some informatics researchers take increasingly formal approaches to terminology development, use and evaluation (see, for example, the paper by Lin and colleagues in the current issue of this journal [17]), others continue to seek to reuse ICD-encoded data for secondary, unintended purposes (see, for example, three papers in the current issue of this journal [18–20]). Some might consider the reliance on such data akin to searching under a streetlamp for ones keys, after losing them in the dark, because the light is better. However, we can hardly blame the researchers. Encoding clinical data is usually a resource-intensive process. Given that most health data are captured in the process of patient care, and that the healthcare entities doing the cap-

turing have limited resources, it follows that the data will be coding only once (if we are lucky), using the method that will yield the greatest immediate benefit. Currently, that method is ICD-encoding, since it is directly tied to required functions such as reimbursement and government reporting. Additional data coding is often seen to be at best an extravagance and at worst a burden. Attention has often been on ways to improve the ICD coding (see, for example, the paper by Prins and Hasman in this issue [18]), rather than on ways to improve the capture of coded data for reuse.

Many researchers have called for change in this status quo (see, for example, the review by Rose and colleagues [21]). Although expressed in a variety of ways, the underlying theme is the same: instead of coding data for a particular purpose, redirect the coding efforts to capture the actual meaning of the data, at the level of detail that is relevant to the clinical context. We can then leave it to computer systems to recode the data, as needed, for specific purposes using specific terminologies.

I believe that we are on the threshold of being able to pursue such a course. Several developments are aligning that favor such a prediction. The decades of terminologic research have yielded a critical mass of results that teach us how to recognize flaws in controlled terminologies and how to correct them. Established standard terminologies, such as SNOMED, LOINC and others, are

expanding in principled ways to cover their respective domains, while new terminologies are being developed with principled approaches to cover additional domains (such as medications).

At the same time, healthcare organizations and government agencies are beginning to acknowledge that the information technology they are depending on to improve healthcare quality and efficiency requires usable clinical (as opposed to administrative) data. Until now, the general approach to acquiring such data has been to add requirements for additional coding of patient care and research information, using additional, single-purpose terminologies. The message that the informatics community has been sending for years is finally getting some attention. For example, the US Committee on Vital and Health Statistics has recommended the use of SNOMED and LOINC for encoding data – not for a specific purpose, but rather so that they can support a variety of "meaningful uses" of electronic health records [22].

There are many challenges to retooling health information systems to capture data with this new breed of standard terminologies. The ability to convert data into forms required for reporting and reimbursement seems likely to be one of the lesser challenges, when compared to tasks such data entry by clinicians and automated understanding of text-based documents. However, the vast body of terminologic research to date supports the hypothesis that the effort will be worthwhile. For example, we will be better able to reuse data when we can employ terminologies that support clinically relevant levels of detail; decision support tools will be easier to create and maintain when they can rely on the classification of terms in the terminology to make their inferences; and those seeking to aggregate population data will be able to do so without concern for time-based semantic drift in the meanings of those data. Such achievements have already been realized at

institutions where informaticians have led the way, using principled, but largely local, terminologies, as has been documented in the 50 years of *Methods of Information in Medicine*. It is fortuitous that as the use of health information technology to actually improve human health becomes mainstream, the high-quality standard terminologies are finally mature enough to be part of the solution.

Acknowledgment and Disclaimer

References

1. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975; 8 (4): 303–320.
2. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982; 307 (8): 468–476.
3. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 1999; 27 (1): 12–17.
4. Cimino JJ. Unpublished data from mail surveys of New Jersey and Florida hospitals; 1979.
5. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005; 40 (5 Pt 2): 1620–1639.
6. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. The New England Journal of Medicine 1985; 312 (25): 1604–1608.
7. Yu AC, Cimino JJ. A Comparison of Two Methods for Retrieving ICD-9-CM data: The Effect of Using an Ontology-based Method for Handling Terminology Changes. J Biomed Inform 2011. (Epub ahead of print)
8. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32 (4): 281–291.
9. Selden C. Unified Medical Language System (UMLS): January 1986 through December 1996 : 280 citations/prepared by Catherine R. Selden, Betsy L. Humphreys. -- Bethesda, Md. (8600 Rockville Pike, Bethesda 20894): US Dept of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, Reference Section ; Pittsburgh, PA. Sold by the Supt of Docs, U.S. G.P.O., 1997. (Current bibliographies in medicine ; 96–98)
10. PubMed search for "UMLS or Unified Medical Language System", performed on February 11, 2011, at http://www.ncbi.nlm.nih.gov/pubmed
11. Methods Inf Med 1995; 34 (1–2): 1–231.
12. Methods Inf Med 1998; 37 (4–5): 311–575.
13. Methods Inf Med 2001; 40 (4): 272–361.
14. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003; 49 (4): 624–633.
15. Benson T (ed). Principles of Health Interoperability HL7 and SNOMED/Edition 1. New York: Springer-Verlag, LLC; 2009.
16. Jiang G, Pathak J, Chute CG. Formalizing ICD coding rules using Formal Concept Analysis. J Biomed Inform 2009; 42 (3): 504–517. Epub 2009 Feb 21.
17. Lin MC, Vreeman DJ, McDonald CJ, Huff SM. A Characterization of Local LOINC Mapping for Laboratory Tests in Three Large Institutions. Methods Inf Med 2011; 50: 105–114.
18. Prins H, Hasman A. Long-term Impact of Physician Encoding on Detail and Number of Recorded Diagnoses Support by a Speciality Specific List of Diseases. Methods Inf Med 2011; 50: 115–123.
19. Trombert Paviot B, Gomez F, Olive F, Polazzi S, Remontet L, Bossard N, Mitton N, Colonna M, Schott A-M. Identifying Prevalent Cases of Breast Cancer in the French Case-mix Databases. Methods Inf Med 2011; 50: 124–130.
20. Kimura M, Nakayasu K, Ohshim T, Fujita S, Nakashima N, Jozaki T, Numano T, Shimizu T, Shimomura M, Sasaki F, Fujiki T, Nakashima T, Toyoda K, Hoshi H, Sakusabe T, Naito Y, Kawaguchi K, Watanabe H, Tani S. SS-MIX: A Ministry Project to Promote Standardized Healthcare Information Exchange. Methods Inf Med 2011; 50: 131–139.
21. Rose JS, Fisch BJ, Hogan WR, Levy B, Marshall P, Thomas DR, Kirkley D. Common medical terminology comes of age, Part Two: Current code and terminology sets – strengths and weaknesses. J Healthc Inf Manag 2001; 15 (3): 319–330.
22. National Committee on Vital and Health Statistics. Report of Hearing on "Meaningful Use" of Health Information Technology. April 28–29, 2009, Washington, DC. http://www.ncvhs.hhs.gov/090518rpt.pdf