This material was originally published in the Journal of the American Medical Informatics Association. Presentation of this material by James J. Cimino is made possible by a limited license grant from the American Medical Informatics Association ("AMIA") which has retained all copyrights in the contribution.

Editorial Comments JAMIA

Controlled Medical Vocabulary Construction: Methods from the Canon Group

The recall and manipulation of medical knowledge and patient information motivates most medical informatics research. Before a computer program can do such sophisticated tasks, the knowledge or information must be represented in the computer. Researchers often see the information representation problem only as a means to an end; dealing with it to the degree necessary to allow them to test different strategies for information retrieval or manipulation. This tendency is reflected in the methods descriptions found in research papers. Although a data model may be presented, there is often little mention made of the vocabulary that is used by the system to represent concepts, or how that vocabulary was developed.

Even in reports that are *about* controlled vocabulary, the *methods* used for its derivation are usually minimal or informal. Virtually nothing has been published in the peer-reviewed literature that describes formal, reproducible methods by which the large standard vocabularies, such as the *International Classification of Diseases* (ICD9)[1] or the proposed Uniform Clinical Data Set (UCDS),[2] were developed or are maintained. Papers reporting research involving the National Library of Medicine's Unified Medical Language System *have* provided careful descriptions of the methods used, but the majority of that work has involved merging and using existing vocabularies, not deriving new vocabularies or expanding existing ones. As a result, comprehensive clinical vocabularies, and the methods for developing them, are in their infancy.[3] This area of work does not yet have a common language of discourse. Is one person's "concept" another's "term"? Does "attribute" mean the same thing as "modifier"?

Given this state of the art, a number of researchers agreed to try to coordinate their individual efforts to move toward a canonical representation of medical information. The group adopted the name Canon Group and a paper outlining their position appears in this issue of the *Journal*.[4] The Canon Group held a retreat in January 1993, to discuss methods for developing and sharing schemes for medical information representation. The first step was to better understand each other's work. Attendees were given a "homework" assignment in preparation for the meeting: using two particular textual chest x-ray reports (agreed on by the group to be generally representative of such reports), encode the findings with some controlled vocabulary and structure of your own choosing, and show the resulting representation in minute detail.

The result was a fascinating collection of presentations that disclosed the terminologic work being done at eight different institutions in a depth not previously reached in conference presentations or research publications. Presenters showed not only their representations of two reports, but also how their vocabularies were developed and why the various aspects of their particular approaches were important to their own applications. The applications spanned the spectrum of medical informatics, including predictive data entry, patient record systems, patient record retrieval (for applications such as outcomes research and quality assurance), natural-language processing, and automated decision support.

To date, only an abstract has been published of this work.[5] In this issue of the *Journal*, a set of papers[6–8] by members of the group re-create the presentation

style of the original retreat: individual models for particular applications, using one or both sample x-ray reports for illustration purposes. Despite the different goals, similarities can be seen in the methods applied in these papers, particularly because having the same x-ray reports in each facilitates direct comparison.

Although these papers include Results sections, they are briefer, and their contents more preliminary, than would normally be found in these pages. The real content is in the Methods sections, where detailed descriptions of the analytic approaches used to *develop* representational schemes can be found, with less emphasis on their application. Each of these papers represents a start at what will ultimately be a long and no doubt painful process. Although the difficulty of the task is recognized, we must start somewhere and start soon. We need formal methods and computer-based tools that can help us with the task. We need research in which controlled-vocabulary development is the focus rather than a stepping stone for work on other theories and applications. The *Journal* is publishing these papers in hopes of stimulating a broader discussion of the issues that surround the representation of medical concepts. Such a discussion could lead to the concerted research effort that will be necessary to move this important area of medical informatics toward maturity.

JAMES J. CIMINO, MD

The Canon group thanks the Digital Equipment Corporation, the IBM Corporation, and Mr. Paul Mongerson, whose generous support made the retreat possible.

References ■

1. United States National Center for Health Statistics. International Classification of Diseases, Ninth Revision, with Clinical Modifications. Washington, DC, Department of Health and Human Services, DHHS 80-1260, 1980.
2. Audet AM, Scott HD. The Uniform Clinical Data Set: an evaluation of the proposed national database for Medicare's Quality Review Program. Ann Intern Med. 1993;119(12):1209–13.
3. United States General Accounting Office. Automated Medical Records: Leadership Needed to Expedite Standards Development. Report to the Chairman/Committee on Governmental Affairs, U.S. Senate, Washington, DC, April 1993. USGAO/IMTEC-93-17.
4. Evans DA, Cimino JJ, The Canon Group. Toward a medical concept representation language. J Am Med Informatics Assoc. 1994;1(3):207–17.
5. Evans DA, Chute CG, Cimino JJ, et al. CANON: towards a medical-concept representation language for electronic medical records (abstract). In: Kahn MG, ed. Proceedings of the 1993 Spring Congress of the American Medical Informatics Association, St. Louis, MO;1993:26.
6. Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. J Am Med Informatics Assoc. 1994;1(3):249–62.
7. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Informatics Assoc. 1994;1(3):218–32.
8. Friedman C, Cimino JJ, Johnson SB. A schema for representing medical language applied to clinical radiology. J Am Med Informatics Assoc. 1994;1(3):233–48.

Correspondence and reprints: James J. Cimino, MD, Center for Medical Informatics, Atchley Pavilion, Room 1310, Columbia-Presbyterian Medical Center, New York, NY 10032.

Received for publication: 12/30/93; accepted for publication: 1/10/94.