ELSEVIER SCIENCE
IRELAND

International Journal of Bio-Medical Computing
34 (1994) 185–194

International Journal of
Bio-Medical
Computing

Data storage and knowledge representation for clinical workstations

James J. Cimino

Center for Medical Informatics, Columbia University, Atchley 1340, Columbia Presbyterian Medical Center, New York, NY 10032, USA

Abstract

The representation of patient information for use in clinical workstations is a complex problem. Ideally, it should be addressed in a way that allows multiple uses of the data, including simple manual review, sharing and pooling across institutions, and as input to knowledge-based decision support systems. To a great extent, this means coding information with controlled medical vocabularies, but it does not mean that all information must be codable before workstations are feasible. This paper defines some of the choices, both current and future, that are available to address the needs of controlled medical vocabularies for representing data and knowledge in clinical workstations and explores some of the implications of those choices.

Key words: Nomenclature; Terminology; Vocabulary; Semantics; Artificial intelligence; Term harmonization

1. Definitions and scope

Desired functions of a clinical workstation include the ability to capture patient information and the ability to use that information, together with appropriate information about patients in general, to assist in the clinical decisions of patient care. In this paper, the author shall refer to the patient-specific information as 'data' and the general, patient-independent information as 'knowledge'. Typical patient data include name, date of birth, weight, physical findings, laboratory results, problem lists, medications and planned procedures. Knowledge includes information about the most common causes of physical findings, the predictive values of tests, diagnostic protocols, drug interactions, and methods for accessing appropriate outside information sources.

The representation of patient data and medical knowledge are major challenges to the development of practical clinical workstations. The representational scheme

SSDI 0020-7101(94)00909-2

used for patient data impacts on the ability to capture, store, retrieve and share that data. The scheme used for knowledge representation impacts on the ability to apply that knowledge to actual patient data. This paper examines some of the issues that influence the choice of data and knowledge representation, reviews the current state of the art, identifies the remaining needs and addresses the four workshop themes of commonality, diversity, utility and barriers.

2. Uses for data and knowledge in the clinical workstation

The patient data found in clinical workstations derives from several sources. Some are recorded at the workstation itself. Some comes from external systems, such as patient registries, financial systems, laboratory systems, discharge summary systems, other workstations, etc. The primary purpose of these data is to assist the clinician with the patient care process. Other purposes include: serving as a legal record, transmitting information to other clinicians, use in billing, inclusion in pooled patient databases for clinical research and quality assurance, and acting as input to knowledge-based systems.

The primary role of knowledge-based systems in the clinical setting is to provide access to general, patient-independent information which is relevant to the decision at hand. In some cases, the knowledge has a passive role, in others it has an active role. Passive knowledge includes information sources such as databases of medical text or bibliographic citations. Searching such a database with a question or, perhaps, some patient data can produce information which must be interpreted by the clinician in order to apply it to a decision. In an active knowledge base, the patient information is processed by the system to identify relevant knowledge which is offered in the form of critiques, alerts, interpretations, etc. Such results can then serve as patient data, in much the same way as the clinician's own interpretations and impressions.

3. Current representation and usage

The representation of data from external systems depends, in large measure, on the scheme used by those systems. Data collected by the clinical workstation itself fall into three general types: observations, interpretations and actions. For each of these types, different data collection methods are used.

Observations include the subjective and objective data, such as the history and physical examination, which are gathered by the clinician during a patient encounter. Such information may be captured through typing by the clinician or through writing or dictation by the clinician with typing by a transcriptionist. This results in free-text data. In many cases, the text can be highly structured, such as with a system-oriented physical examination. In other cases, the data can be captured through the use of controlled vocabulary, with or without the addition of free text. Such data collection is done routinely with encounter forms, such as in the COSTAR system [1]. Experimental work has been done to allow the clinician to enter data directly through the use of controlled vocabulary, using touch screen [2], mouse pointer [3], and voice recognition [4]. In limited domains (such as mammography reports) where the vocabulary is fairly restricted, successful commercial systems using voice recognition have been developed.

Interpretations are captured in a similar manner to observations. However, because they represent a synthesis and crystallization of the observation data, there is often more of an effort to obtain such data as coded information. Vocabularies such as ICD9-CM and SNOMED have been used for this purpose [5,6]. Like interpretations, it is often desirable to record actions using a controlled vocabulary. Typical actions include the ordering of diagnostic or therapeutic procedures and the prescription of medications. When such data are recorded using controlled terminology, they can be used as input to other systems, such as order entry, pharmacy and laboratory systems.

The way in which data are organized and stored are as important to usage as the way in which they are represented. Free text may be the easiest to store but may the most difficult to use. For example, storing 'blood sugar was 117' is certainly readable to a human reviewing the information but is less useful for summary reporting of all recent blood sugars. Usefulness can be increased if the text can be organized chronologically and structured by context (for instance, to display the heart examination over time). Also useful for retrieval purposes are various indexing schemes [7,8]. Coding the information can increase the usefulness of the text by improving the relevance and recall of retrievals [9]. So, for example, if blood sugar results are stored as a code (say, '651234' for 'blood sugar') and a result (say, '117'), producing a graph of all results tied to test 651234 is relatively straightforward.

Passive knowledge is usually stored as narrative text (such as journal article abstracts) or as indexed text (such as MEDLINE citations indexed with MeSH). Active knowledge, on the other hand, must be coded in order to be manipulated by the expert system or decision support system that uses it. When such a system takes patient data as input, that data must be converted to a coded form which is usable by the system, if it is not already stored this way. This task is usually carried out by the clinician; however, a number of efforts, particularly those using the Unified Medical Language System (UMLS) [10], are working to develop ways to automate this process.

4. Current needs

There is clearly a pressing need for improvement in user interfaces to allow capture of clinical data in all forms; however, this subject is to be dealt with elsewhere in the workshop. Assuming that appropriate user interfaces can be constructed, there remain a number of data and knowledge representation issues to be resolved, including: how to store the data, how much of the data to code and what codes to use.

Some clinical data take the form of traditional data types, such as integers, character strings and dates. In order to have meaning, such information must have a context. Sometimes the context is provided by the database design and sometimes it is provided by pairing it with codes. Fig. 1 depicts several ways in which blood sugar results can be encoded and stored in a relational database. Depending on the scheme, the context is indicated by the table, by the row in a table or by a tupple

(A)

Blood Sugars

Value
100
121

(B)

Test Results

Blood Sugar
100
121

(C)

Test Results

Test	Value
"Blood Sugar"	100
"Blood Sugar"	121

(D)

Test Results

Test	Value
651234	100
651234	121

Fig. 1. Four ways to code and store blood sugar results. In (A), the 'table-oriented' approach, there is a table called 'Blood Sugars' with a column called 'Value', where each contains the test results. In (B), the 'column-oriented' approach, the table is used for all test results with a column for each test; blood sugar results appear as values in the column called 'Blood Sugar'. In (C), the 'uncoded, row-oriented' approach, the table is used for all test results and, rather than having test-specific columns, one column is used to hold the name of the test and one is used to hold the result. Finally, (D) the 'coded, row-oriented' approach, is similar to (C) except that blood sugar tests are stored by code rather than by name. Note that although a relational model is used here, other storage schemes could be used with equal diversity.

in a table. In each case, the amount of information is identical. In the first two schemes (table-oriented and column-oriented), the fact that blood sugars are to be stored is taken into account in the database design. In the latter two schemes (uncoded, row-oriented and coded, row-oriented), blood sugar results are accommodated by being stored in generic columns which can accommodate many different types of test results. The test type can be stored as data of type free text or as coded information. The actual test results are stored, in each case, as data of type integer.

Coded data are stored through the use of a set of legal codes, where each code corresponds to some meaning in a controlled vocabulary. The vocabulary may be quite small, such as 'Gender' (e.g., 'M' means 'male' and 'F' means 'female'), or quite large, such as 'Diagnosis' (using, for example, ICD9 codes). These controlled vocabularies are a crucial aspect of clinical workstations because they provide the link to automated functions, such as data summaries and decision support. There are presently some controlled vocabularies which can be used for recording clinical data for use in workstations, but there is no single vocabulary, or even set of vocabularies, which is accepted as providing coverage for all desired data.

The tension between narrative text and coded data is likely to exist for some time. There has been work directed at encoding all clinical data [11,12]; however, caution is warranted. It is not at all clear that, just because data is encoded, it is more useful. The large terminologies needed to encode 'all' medical narrative may be too cumbersome for use with knowledge-based systems. Such systems focus on encoding only small, relevant portions of the clinical record, using a small, closely-managed terminology. An alternative to attempting complete coding, therefore, is to work toward coding those data which are most useful in coded form and leaving the remainder as narrative text. The narrative is still useful for human review, while the portion being coded can be increased over time, as our ability to use such coded data increases (in more sophisticated decision support systems, for example).

5. Commonality

Clinical workstations will be developed by different groups for use by different types of clinicians. What they all will have in common, however, is that they will deal with data about human patients and about medical knowledge to be used in caring for those patients. If workstations made use of the same conventions for representing data and knowledge, a number of benefits would be possible, including the ability to share data and knowledge and to work in cooperation to build the large controlled vocabularies needed for coding.

First, at the most basic level, there should be a common set of data types which are recognized by all workstations. Some standards have been developed which are applicable to this requirement. For example, the Arden Syntax for medical logic modules includes a number of primitive data types for use in representing patient data for manipulation by medical logic [13].

Second, although widespread agreement on a single controlled vocabulary for use in clinical workstations is not likely to occur soon, there may be areas of commonality which can be developed. One is the general classes of concepts which are codable. General classes include, for example, 'complaints', 'physical findings', 'medications', 'diagnostic tests' and 'problems'. Such classes can be found in the axes of SNOMED, the trees of MeSH and the Semantic Types of the UMLS.

Third, a general syntax should be created for representing controlled vocabularies. For a given term in a given controlled vocabulary, it should be possible to identify some very basic information about the term, such as its name, the class it is in, alternative names and relationships to other terms in the controlled vocabulary. Standards such as ASN.1 may be useful for this purpose.

Fourth, given common vocabulary classes and representational schemes, some progress should be possible toward the creation of common content for some subsets of the controlled vocabulary. For example, enumeration of all drugs made by all drug companies should be possible and codable (in fact, there are a number of coding systems which may already offer a solution to this particular problem). Efforts such as the UMLS and Galen [14] could provide a central repository where such terminology subsets could be formed, maintained and distributed.

Fifth, once some commonality is achieved with patient data representation, common methods for representing medical knowledge can be developed. For example, the development of the Arden Syntax was possible because of commonality among certain decision support systems and their underlying clinical information systems. Besides Arden, which allows representation of medical logic, the UMLS Information Sources Map [15] provides a format for representing knowledge about passive and active knowledge sources. Common representational schemes for other types of knowledge should be possible, such as for disease finding patterns used in diagnostic programs (e.g., QMR [16], DXplain [17] and Iliad [18]).

6. Diversity

The creation of clinical workstations today will, of necessity, involve a great deal of diversity. In some cases, the content needed for representing patient data will be site-specific and therefore, by definition, diverse. In addition, each site will include its own particular applications which will impose their own demands on data structure and storage.

Although there are many potential areas for the development of commonality, waiting for their creation would cause unnecessary delay in the development of clinical workstations. As workstations are developed and sharing of patient data, vocabularies, knowledge and representational structures begins to occur, additional areas of commonality can be identified and developed. Fig. 2 depicts the way in which evolution can take place, from unshared, unstructured data to common, coded information. Initially, most data are represented as free text, some of which is structured. A minority of the data are coded and, where coding exists, local coding schemes are used. With the development of standards, the borders between different types of data will shift in favor of those which are structured, coded and sharable.

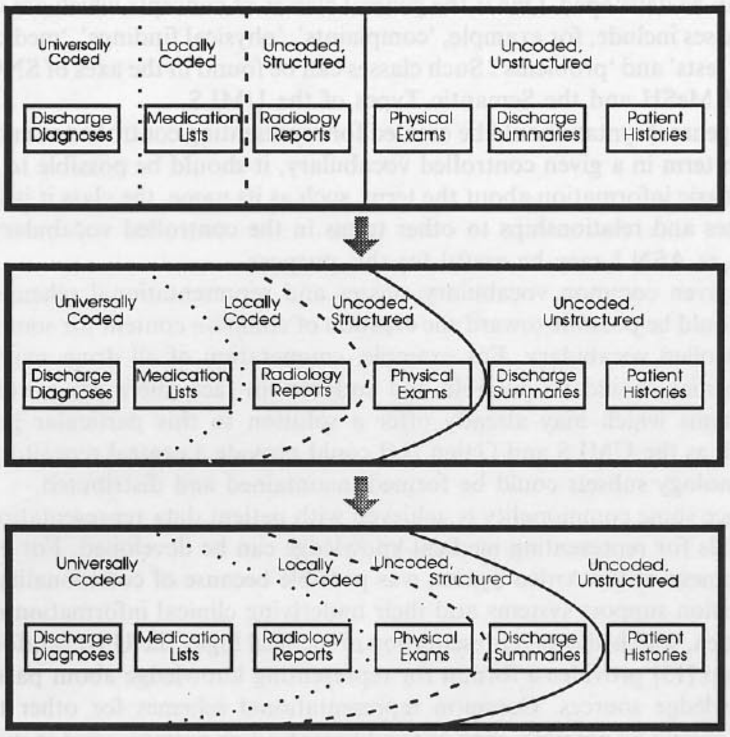

Fig. 2. An orderly transition to more coded patient data. The top panel depicts the present state of patient data coding, with most data completely unstructured (such as narrative text). Succeeding panels give examples of how various data can move to become fully coded and sharable. For example, radiology reports may have some structure in present systems but current efforts are under way to provide a coding scheme for such data; should that scheme find universal acceptance, it could join information sets, such as discharge diagnoses (coded in ICD9) which can be shared today for pooling across institutions or used by decision support tools.

7. Utility

As we 'push the envelope' shown in Fig. 2, the functionality of the clinical workstation will increase. Our success can be measured by the performance of capabilities such as sharing patient data, sharing clinical applications, cooperating to expand coded data, and sharing medical knowledge.

With a common representational scheme for patient data, sharing of that data offers the potential for improving patient care, particularly with respect to continuity aspects. For example, if a patient's medication list in some out-patient system is available when a patient is admitted to the hospital, costly mistakes can be avoided, such as stopping a necessary medication or adding a new one that interacts with an existing one. Patient care can also be made more efficient through the sharing of data, since time, money and risk can be saved by obviating the need to collect data that has been obtained elsewhere. For example, if a patient has recently undergone bloodwork and X-ray examinations, that information may be used in lieu of repeating the tests.

Another benefit of sharable patient data will be the ability to carry out clinical research across institutions. This will be particularly useful for outcomes research. For example, following patients with a specific condition who undergo a variety of surgical and medical therapies can allow comparisons of relative effectiveness of each treatment.

Clinical workstations are more than data recording and display devices; they include a variety of applications for processing that data to assist in patient care. These applications are often 'home grown' because of the complexity of retrieving and using patient information. Turn-key systems are often difficult to install, not for technical reasons, but because they must be adapted to work with each institution's patient data format and content. As standard representational schemes are developed for the data, clinical applications will become more sharable. For example, a prescription-writing program that makes use of a standard database of medications could be written in one institution and used by many others with a minimum of effort.

Once institutions begin to share data, there will be pressure to expand the portion of the data that can be coded through controlled vocabularies. With a common representational scheme for those vocabularies, cooperative efforts can be undertaken to expand the vocabularies. For example, one coalition of researchers has developed a common representational scheme for X-ray findings and has had early success with the construction of a common, sharable lexicon of actual findings [19].

Finally, the ability to share medical knowledge will be facilitated not just by a common knowledge representation scheme but by a common data representation as well. For example, a great deal of medical knowledge can be expressed as 'don't do X to patients with condition Y'. The Arden Syntax was developed to allow the expression of such rules. When an author of a medical logic module writes such a rule for her own institution, she knows what she means by X and Y and how they are represented in her system. But when someone at another institution attempts to include the rule in his system, he may not find that X and Y are so readily obtained. In fact, the biggest impediment to sharing such knowledge in Arden medical logic modules appears to be not the ability to write them, transmit them or interpret/compile them, but the ability of the receiving institution to provide the appropriate patient data (Hripcsak, pers. commun.). A common scheme for representing patient data will permit the sharing of retrieval strategies — the so called 'curly bracket' information in Arden module data slots.

As we become more successful in representing, capturing and storing patient data, we become more able to apply decision support techniques in automated ways to provide clinicians with appropriate knowledge, of both the passive and active forms, when needed. For example, because patient problems and procedures are coded with ICD9, one workstation is able (using the UMLS) to use patient data to help automate the retrieval of medical knowledge through literature searches [20].

8. Barriers

The single greatest barrier to the development of clinical workstations may be the need for appropriate controlled vocabularies. Creation of these vocabularies requires a mechanism for sharing which conveys structure, content and meaning. It is this last aspect which is the most troublesome. For years, controlled vocabularies have been shared without explicit inclusion of the meanings of terms. The absence of the meanings may explain the lack of widespread acceptance and use of these vocabularies. For example, when should a term such as 'Anemia Not Elsewhere Classified' be used? The actual meaning of such a term is based on the exclusion of all other terms in the vocabulary. New methods for vocabulary representation are clearly needed before we can begin to cooperate on sharing them and the data coded with them [21].

A number of workers have proposed that controlled vocabularies include structured, coded information about the terms themselves [22–24]. The use of this approach to representation may offer benefits to clinical workstations beyond the ability to encode data. For example, consider a controlled vocabulary for representing laboratory tests. What structured, coded information would be included in such a vocabulary that would provide information about the meaning of the terms? Certainly, there would have to be information about what specimens are used for tests and what the tests measure. This type of representation is applicable to more than just term definitions. It reflects the underlying model of clinical laboratory data and is also relevant to medical knowledge about laboratory results.

The selection of a representation scheme is one task needed to overcome the barrier of controlled vocabulary. Another task is the development of content. Content will need to be sufficiently extensive to provide for domain completeness, with inclusion of multiple recognized synonyms for terms, while avoiding redundancy. The organization of the vocabularies will need to recognize the coexistence of multiple legitimate classification schemes. Criteria such as these have heretofore eluded large controlled vocabularies [25].

9. Future directions

The controlled vocabulary barrier is a substantial one. Fortunately, a number of efforts are under way to address this issue. In Europe, CEN Technical Committee 251 is striving for a representational scheme that can permit sharing of controlled vocabularies across political and linguistic borders. In cooperation with that committee, the Advanced Informatics in Medicine program, also in Europe, is sponsoring the Galen project to develop content where none currently exists.

In the USA, the National Library of Medicine's UMLS project has developed, and continues to develop, a common representational scheme for terms from disparate vocabularies. The Computer-based Patient Record Institute, a consortium of health care, vendor and professional organizations is studying the UMLS and other existing systems for providing the coded data needed for a electronic medical records and clinical workstations.

More work must be done to elucidate the precise challenges facing the data and knowledge representation needs for clinical workstations. Those who have worked on the problem are acutely aware of the present absence of satisfactory solutions. The answer is not simply to enlarge some already-large list of terms such that it can cover every possible medical utterance. Instead, thought must be given to the creation of terminologies that at the same time accommodate the rich language of medical information while being constructed to facilitate storage and retrieval of the information for sharing with others and as fodder for the automated application of medical knowledge to clinical problem solving. A daunting task. But, just as we must walk before we can run, so should we attempt an incremental approach to this task. We can learn and build on the steps that have already been taken as we continue to move forward.

10. References

1 Beaman PD, Justice NS and Barnett GO: A medical information system and data language for ambulatory practice, *Computer*, (November 1979) 9–17.

2 Greenes RA, Barnett GO, Klien SW, Robbins A and Prior RE: Recording, retrieval and review of medical data by physician-computer interaction, *N Engl J Med*, 282 (1970) 307–315.

3 Cimino JJ and Barnett GO: The physician's workstation: recording a physical examination using a controlled vocabulary. In: *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care, 1–4 November 1987, Washington* (Ed: W.W. Stead), IEEE Computer Society Press, Los Angeles, 1987, pp. 287–290.

4 Wulfman CE, Rua M, Lane CD, Shortliffe EH and Fagan LM: Graphical access to medical expert systems: V. Integration with continuous-speech recognition, *Methods Inf Med*, 32 (1993) 33–46.

5 Cimino JJ, Barrows RC and Allen B: Adapting ICD9-CM for clinical decision support (abstract). In: *Proceedings of the Second Annual Educational and Research Conference of the American Medical Informatics Association. 1992 May 6–9, Portland (OR)* (Ed: MA Musen), The American Medical Informatics Association, Bethesda, MD, 1992, p. 34.

6 Satomura Y and Amaral MB: Automated diagnostic indexing by natural language processing, *Med Inform*, 17(3) (1992) 149–163.

7 Safran C, Porter D, Lightfoot J, Rury CD, Underhill LH, Bleich HL and Slack WV: ClinQuery: a system for online searching of data in a teaching hospital, *Ann Intern Med*, 111(9) (1989) 751–756.

8 Vries JK, Marshalek B, D'Abarno JC, Yount RJ and Dunner LL: An automated indexing system utilizing semantic net expansion, *Comput Biomed Res*, 25(2) (1992) 153–167.

9 Payne TH, Goroll AH, Morgan M and Barnett GO: Conducting a matched-pairs historical cohort study with a computer-based ambulatory medical record system, *Comput Biomed Res*, 23 (1990) 455–472.

10 Lindberg DAB and Humphreys BL: Toward a unified medical language system. In: *Medical Infor-*

matics Europe '89; Proceedings of the Seventh International Congress; 1987 September 11–15, Rome Springer-Verlag, Berlin, 1987, pp. 123–131.

11 Gabrieli ER: A new electronic medical nomenclature, *J Med Syst*, 13(6) (1989) 355–373.

12 Cote RA and Rothwell DJ: The classification-nomenclature issues in medicine: a return to natural language, *Med Inform*, 14(1) (1989) 25–41.

13 American Society for Testing and Materials: Specification for defining and sharing modular health knowledge bases (Arden Syntax for medical logic systems). In: *Annual Book of ASTM Standandards*, Vol. 14.01, ASTM, 1992, pp. 539–587.

14 Rector AL, Nowlan WA and Kay S: Conceptual knowledge: the core of medical information systems. In: *MEDINFO '92: Proceedings of the 7th World Congress on Medical Informatics* (Eds: KC Lun, P Degoulet, TE Piemme and O Reinhoff), North-Holland, New York, 1992, pp. 1420–1426.

15 Lindberg DAB and Humphreys BL: The UMLS knowledge sources: tools for building better user interfaces. In: *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Washington DC, November, 1990* (Ed: RA Miller), 1990, pp. 121–125.

16 Miller RA and Masarie F: Quick Medical Reference (QMR): A microcomputer-based diagnostic decision-support system for general internal medicine. In: *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Washington DC, November, 1990* (Ed: RA Miller), 1990, pp. 986–998.

17 Barnett GO, Cimino JJ, Hupp JA and Hoffer EP: DXplain — an evolving diagnostic decision-support system, *J Am Med Assoc*, 258(1) (1987) 67–74.

18 Warner HR, Haug P, Bouhaddou O, Lincoln M, Warner Jr H, Sorenson D, Williamson JW and Chinli F: ILIAD as a expert consultant to teach differential diagnosis. In: *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care* (Ed: RA Greenes), IEEE Computer Society Press, New York, 1988 pp. 371–376.

19 Evans DA, Chute CG, Cimino JJ, Greenes RA, Hersh WR, Huff SM and Musen MA: CANON: towards a medical-concept representation language for electronic medical records (abstract). In: *Proceedings of the Third Annual Educational and Research Conference of the American Medical Informatics Association, 10–12 May 1993, St. Louis, MO* (Ed: MG Kahn), The American Medical Informatics Association, Bethesda, MD, 1993, in press.

20 Cimino JJ, Johnson SB, Aguirre A, Roderer N and Clayton PD: The Medline button. In: *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care, 8–10 November 1992, Baltimore, MD* (Ed: ME Frisse), McGraw-Hill, New York, 1992, pp. 81–85.

21 Masys DR: Of codes and keywords: standards for biomedical nomenclature, *Acad Med*, 65 (1990) 627–629.

22 Rector AL, Nowlan WA and Kay S: Unifying medical information using an architecture based on descriptions. In: *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Washington DC, November, 1990* (Ed: RA Miller), 1990, pp. 190–194.

23 Rossi Mori A, Galeazzi E, Gangemi A, Pisanelli DM and Thornton AM: Semantic standards for the representation of medical records, *Med Decision Making*, (Suppl) 11 (1991) S76–S80.

24 Cimino JJ: Saying what you mean and meaning what you say: coupling biomedical terminology and knowledge, *Acad Med*, 68(4) (1993) 257–260.

25 Cimino JJ, Hripcsak G, Johnson SB and Clayton PD: Designing an introspective, controlled medical vocabulary. In: *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, 5–8 November 1989, Washington* (Ed: LW Kingsland), IEEE Computer Society Press, Washington, 1989, pp. 513–518.