# THE PROBLEM OF CODIFICATION

by Prof. Michele Crudele of "Campus Bio-Medico" University of Rome

December 10, 1998 (links reviewed on November 3, 2007)

## *1. Introduction*

The problem of how to code or classify pathologies has been addressed for many years. Initially it was useful for statistical purposes (e.g. causes of death) in epidemiological studies. Later on, while the interchange of patient data among different medical specialists was increasing, the need to identify pathologies in a non-ambiguous way grew substantially.

Given the wideness and the difficulty of organizing all the knowledge regarding pathologies, it is not surprising that there is not a universal classification system, worldwide accepted and used. Moreover, different specialities have developed their own coding system for specific purposes. The World Health Organization has a role in defining standards, but they must be accepted at a national level by all the countries. Other institutions are working at different levels (medical societies, national boards, international committees) with different approaches.

The process of establishing a coding system is mainly composed by two parts: the nomenclature and the classification. Giving a definite, non ambiguous name to the pathology is the task of a nomenclature. Assigning a structured numerical code (e.g. following a tree structure) is the task of a classification. By using the code number, a strict relation between different languages and also different coding systems can be established.

## *2. An overview of the most important coding systems*

Here you can find an alphabetical list, with a brief and simplified explanation, of the most important codification systems in use at the beginning of 1998. It is limited to the general purpose classification systems for pathologies. A more detailed description follows.

### ACR Index

ACR Index is the most widely used system in radiological departments. It is not included in UMLS 1998 database, but may be included in 2000.

### ICD-9

The International Classification of Diseases, 9th revision, issued by the WHO, used mainly for statistical purposes.

### ICD-9-CM

ICD-9-CM is the Clinical Modifications of the ICD-9, comprising also procedures. Widely used for clinical, administrative and statistical purposes.

### ICD-10

ICD-10 is the latest version of the International Classification of Diseases. Not yet adopted by many Countries.

### ICD-10-CM

ICD-10-CM will be the Clinical Modifications to the ICD-10.

### MeSH

The Medical Subject Headings comprise the National Library of Medicine's (USA) controlled vocabulary used for indexing articles, for cataloguing books and other holdings. Its main use is in MEDLINE (biomedical literature) database. Is it very wide, but not restricted to pathologies. The numbering scheme is rarely used.

### THE READ CODES

The Read Codes are a U.K. developed comprehensive list of terms intended for use by all healthcare professionals to describe the care and treatment of their patients.

### SNOMED

A comprehensive, multiaxial indexing system of the entire medical record, including signs and symptoms, diagnoses, and procedures. SNOMED International -either the complete dataset or the Microglossary for Pathology- has been or is being translated into Chinese, French, Danish, Hungarian, Italian, Japanese, Portuguese, Russian, Spanish, and Turkish.

### UMLS

A very wide project by the National Library of Medicine (USA), including a metathesaurus of different coding systems, with cross relations among them. It also serves as a basis for research on natural language retrieving systems.

## 2. A detailed description of the most important coding systems

### ACR Index

The American College of Radiologists (ACR) recommends a system for diagnostic classification based on a tree structure divided in two main branches: an anatomy code and a pathology code.

The *anatomy code* has 2 to 4 digits, before the decimal point
The *pathology code* has 2 to 5 digits, after the decimal point

Example:

Parosteal sarcoma of popliteal surface of femur
4. Indicates skeletal system
45. Indicates knee and leg
451. Indicates distal end of femur

> .3 Indicates neoplasm or neplastic-like condition
> .32 Indicates primary malignant neoplasm of bone
> .322 Indicates neoplasm of osseus origin
> .3222 Indicates parosteal osteosarcoma

451.3222 Parosteal osteosarcoma of distal end of femur

A reduced list (with only 1 digit anatomy) of the ACR Index can be browsed at http://www.xray.hmc.psu.edu/acr_codes/acr_codes.html *(no longer available in 2007; see now www.intrad.ch/intrad/**acr**/)*

A site using ACR codes for searching through radiology cases (for educational purposes) is http://johns.largnet.uwo.ca/med/i-way.html *(no longer available in 2007)*

The official site for ACR is www.acr.org. It does not contain information about the ACR Index.

The cost of the electronic version of the ACR Index for members of the ACR is U.S. $ 82.50 (non-members: U.S. $ 157.50)

### ICD-9-CM

The International Classification of Diseases, Ninth Revision, Clinical Modification, is a classification system that groups related disease entities and procedures for the reporting of statistical information. The clinical modification of the ICD-9 was developed by the National Center for Health Statistics for use in the United States.

The ICD-9-CM consists of :

- a tabular list containing a numerical list of the disease code numbers in tabular form,
- an alphabetical index to the disease entries,
- and the classification system for surgical, diagnostic, and therapeutic procedures (this last section is not part of the international version of ICD)

ICD-9-CM is used for:

- Classifying morbidity and mortality information for statistical purposes
- Indexing of hospital records by disease and operations
- Reporting of diagnosis by physicians
- Data storage and retrieval
- Tool in reporting national morbidity and mortality data
- Basis of DRG (Diagnostic Related Groups) for hospital reimbursement
- Reporting and compiling of health care data to assist in
    - evaluating the appropriateness and timeliness of medical care
    - planning health care delivery systems
    - determining patterns of care among health care providers
    - analyzing payments for health services
    - conducting epidemiological and clinical research

ICD-9-CM Maintenance:

Responsibility for maintenance of the classification system is shared between the National Center for Health Statistics, or NCHS (diagnosis classification) and the Health Care Financing Administration, or HCFA (procedure classification). Suggestions for modifications come from both the public and private sectors and are submitted prior to a scheduled meeting. The meetings are open to the public and comments are encouraged at the meeting and in writing. No decisions are made at the meetings. The committee's role is advisory. All final decisions are made by the Director of NCHS and the Administrator of HCFA.

More information can be found in the official site http://www.icd-9-cm.org/.

A complete table of the ICD-9-CM codes is available at http://www.mcis.duke.edu/standards/termcode/icd9/1tablar.html *(no longer available in 2007; see now http://icd9cm.chrisendres.com/index.php)*

The use of ICD-9-CM code is free of charge.

In Italy ICD-9 is compulsory for death records, while ICD-9-CM is compulsory for compiling medical records in order to calculate the DRG (Diagnosis Related Groups) for reimbursement. It it used by physicians in hospitals, but not always by radiologists.

## *ICD-10*

The Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems is the latest in a series that was formalized in 1893 as the Bertillon Classification or International List of Causes of Death. While the title has been amended to make clearer the content and purpose and to reflect the progressive extension of the scope of the classification beyond diseases and injuries, the familiar abbreviation "ICD" has been retained. In the updated classification, conditions have been grouped in a way that was felt to be **most suitable for general epidemiological purposes and the evaluation of health care**. A version based on ICD-10 (ICD-10-CM) is in preparation.

## *MeSH*

The Medical Subject Headings comprise NLM's controlled vocabulary used for indexing articles, for cataloguing books and other holdings, and for searching MeSH-indexed databases, including MEDLINE.

MeSH terminology provides a consistent way to retrieve information that may use **different terminology for the same concepts**.

MeSH organizes its concepts in a hierarchical structure available at http://www.nlm.nih.gov/mesh/meshhome.html. Therefore searches for a broad concept may include articles indexed to narrower concepts. The first level is divided into:

A. Anatomy
B. Organisms
C. Diseases
D. Chemicals and Drugs
E. Analytical, Diagnostic and Therapeutic Techniques and Equipment
F. Psychiatry and Psychology
G. Biological Sciences
H. Physical Sciences
I. Anthropology, Education, Sociology and Social Phenomena
J. Technology and Food and Beverages
K. Humanities
L. Information Science
M. Persons
N. Health Care
Z. Geographical Locations

The Category C *Diseases* is composed by

C1. Bacterial Infections and Mycoses
C2. Virus Diseases
C3. Parasitic Diseases
C4. Neoplasms
C5. Musculoskeletal Diseases
C6. Digestive System Diseases
C7. Stomatognathic Diseases
C8. Respiratory Tract Diseases
C9. Otorhinolaryngologic Diseases
C10. Nervous System Diseases
C11. Eye Diseases

C12. Urologic and Male Genital Diseases
C13. Female Genital Diseases and Pregnancy Complications
C14. Cardiovascular Diseases
C15. Hemic and Lymphatic Diseases
C16. Neonatal Diseases and Abnormalities
C17. Skin and Connective Tissue Diseases
C18. Nutritional and Metabolic Diseases
C19. Endocrine Diseases
C20. Immunologic Diseases
C21. Injuries, Poisonings, and Occupational Diseases
C22. Animal Diseases
C23. Symptoms and General Pathology

In normal use, lower levels are not normally coded with letters and/or numbers, which makes sometimes difficult to cross-reference with other systems. Anyway, the full inclusion of the MeSH terms in the UMLS project provides sufficient relations with other codes.

The MeSH vocabulary is continually updated by subject specialists in various areas. Each year hundreds of new concepts are added and thousands of modifications are made. 1998 MeSH includes more than 18,000 main concepts and over 80,000 cross-references.

For more detailed information concerning MeSH see the MeSH Fact Sheet.

A search engine for MeSH terms is available at http://www.nlm.nih.gov/mesh/MBrowser.html

The use of MeSH terms is free. A complete set of the MeSH can be found at http://www.nlm.nih.gov/mesh/filelist.html provided you agree with the Memorandum of understanding.

## *THE READ CODES*

The Read Codes (British Crown Copyright) are a comprehensive list of terms intended for use by all healthcare professionals to describe the care and treatment of their patients. They enable the capture and retrieval of patient centered information in natural clinical language within computer systems.

The Read Codes cover such topics as occupations, signs and symptoms, investigations, diagnoses, treatments and therapies, drugs and appliances and much more. This enables the recording within the computer system of anything from a summary of the episode of care to potentially a full electronic patient record if desired.

Version 3 is a U.K. thesaurus of clinical terms which enables the creation of an individual's clinical record within person based systems. This allows the meanings of common clinical terms to be shared, thus facilitating the communication of patient centred records between appropriate healthcare professionals.

The Version 3 file structure uses the actual Read Code simply as a label for the term. The hierarchy position is determined by simple electronic relational tables. This allows an infinite number of levels of detail and allows codes and their terms to be moved to form a hierarchical structure which reflects current clinical thinking. The file structure also allows these terms to be augmented by qualifiers.

The thesaurus was developed in full consultation with the various clinical professions, their Royal Colleges and Associations. Three major terms projects have covered Clinical (medical), Professions Allied to Medicine, Nursing, Midwifery and Health Visiting terms. Over 2,000 clinicians have been involved and these terms and qualifiers are currently undergoing refinement at partnership sites. In addition considerable work has been carried out to enhance the drug and appliances dictionary.

Read Code terms are provided, where appropriate, with validated cross-references to ICD-9 and ICD-10. The cross-reference is presented either as a single code or, where a number of alternatives exist, the most likely cross-reference is highlighted as the default, with an opportunity for refinement within the target classification, if applicable.

Normal licences apply only to UK. The free of charge Standard Licence 1 (Evaluation) enables assessment of the depth and breadth of the Read Codes through a demonstration program. No clinical or development use is permitted and no updates are provided.

The official site for the Read Codes is http://www.cams.co.uk/index.htm

## *SNOMED*

SNOMED is the Systematized Nomenclature of Human and Veterinary Medicine. SNOMED International was introduced in 1993 and is traceable to its roots in the early 1960s as the Systematized Nomenclature for Pathology. SNOMED International is a comprehensive, **multiaxial** nomenclature classification work created for the indexing of the entire medical record, including signs and symptoms, diagnoses, and procedures. Its unique design will allow full integration of all medical information in the electronic medical record into a single data structure. The most recent version of SNOMED International (ver. 3.4) contains more than 150,000 terms and term codes in 11 separate modules. The starting letter, followed by a hyphen, identifies the module.

**T- Topography** A functional anatomy for human and veterinary medicine.
**M- Morphology** Terms used to name and describe structural changes in disease and abnormal development.
**F- Function** Terms used to describe the physiology and pathophysiology of disease processes.
**L- Living Organisms** Living organisms of etiological significance in human and animal disease.
**C- Chemicals, Drugs, and Biological Products** Including pharmaceutical manufacturers.
**A- Physical Agents, Activities, and Forces** A compilation of physical activities, physical hazards, and the forces of nature.
**J- Occupations** Developed by, and used with permission from, the International Labour Office in Geneva, Switzerland.
**S- Social Context** Social conditions and relationships of importance to medicine.
**D- Diseases/Diagnoses** A classification of the recognized clinical conditions encountered in human and veterinary medicine.
**P- Procedures** A classification of health care procedures
**G- General Linkages/Modifiers** Linkages, descriptors, and qualifiers to link or modify terms from each module
**X- List of pharmaceutical companies**

The Diseases/Diagnoses Module is a compilation of disorders that include essentially all of the named diseases, diagnoses, and syndromes used in human and veterinary medicine. These disorders are placed into classes and are arranged either by organ systems or by their underlying etiology.

Ten of sixteen sections found in this module classify disorders in the organ system format that parallels the other SNOMED modules. The remaining six sections categorize disorders as either congenital, metabolic, and nutritional, occurring in the pregnancy and perinatal period: injury and poisoning; infectious; or by victim status. Together they form a coherent classification of all the diseases, diagnoses, and syndromes encountered in medicine. Each term is placed in a hierarchy of related terms, thus all gastrointestinal and all cardiovascular disorders are placed into their respective sections. This, however, does not commit these terms to a single hierarchy of disorders.

For example, although tuberculosis of the pleura is assigned a code number in the infectious diseases chapter - DE-14820 - it is cross-referenced to the topography hierarchy - T-29000. The code DE-14820 carries the English terms and not only the references to T-29000, *Pleura*, but also to L-21801, *Mycobacterium tuberculosis hominis*, with synonyms *human tubercle bacillus* and *Koch's bacillus*.

The double hierarchy now becomes evident because one can search all diseases of the pleura in the chapter on respiratory system diseases carrying the cross-reference code to pleura, T-29000. Cross references are in effect a mechanism for producing polyhierarchies specifying in detail what additional relationships are known about a disorder and placing terms in more than one hierarchy.

Placement of terms into multiple hierarchies is a feature of particular important for search and retrieval and is a powerful mechanism for knowledge representation. In the current edition of SNOMED essentially all of the disorders listed in the ICD-9-CM have been isolated, individually listed, and each assigned a specific SNOMED termcode.

Two examples are taken from the Chapter on Injuries:

| **SNOMED** | **Description and related axis** | | | | **ICD-9-CM** |
|---|---|---|---|---|---|
| DD-33--- | Open Wounds of the Limbs | | | | |
| DD-33620 | Open wound | of knee | without | complication | 891.0 |
| | M-14010 | T-D9200 | G-C009 | F-01450 | |
| DD-33621 | Open wound | of knee | with | complication | 891.1 |
| | M-14010 | T-D9200 | G-C008 | F-01450 | |

In this example, if one were to code without the G axis, which adds context, we would code the knee, the open wound and the complication only in DD-33621. In DD-33620, if there is no complication, then why code it? The physician did not observe any complication and wants this negative finding recorded. Therefore, there is a necessity to code negative as well as positive findings and for this we need modifiers and links as found in the G axis.

Between the two diagnoses used as examples, the only difference is the "without" (G-C009) and the "with" (G-C008). SNOMED coding of complex diagnoses allows the capture of medical information decomposed into basic concepts tied together by G-axis links which preserve the context for retrieval.

A reduced version called SNOMED Microglossary for Pathology is available.

SNOMED International is being accepted worldwide as a standard for indexing medical record information. In addition, SNOMED is specified as the controlled terminology and message standard for interchange of biomedical images and image-related information in the DICOM (Digital Imaging and Communications in Medicine) standards. The SNOMED DICOM Microglossary (SDM) provides a context-sensitive controlled terminology for the clinical specialties that perform or depend upon diagnostic imaging procedures. Content in the SNOMED DICOM Microglosssary is mapped from SNOMED, LOINC, and other clinical nomenclature systems. The SDM contains four types of database tables:

1. A Context Group Summary Table which describes each context group in the SDM
2. Many Context Group Tables, each of which provide context-dependent value sets for a given DICOM data element
3. A Template Summary Table, which describes each template in the SDM
4. Many Template Group Tables, each of which specifies the set of properties that fully describe a given concept.

Supplement 23 of the DICOM Standard defines the structured reporting specifications for the transfer of observation data accompanying an image. This data includes text descriptions of significant findings, measurements, annotation/labelling, etc. These specifications would enable an ultrasound technologist, for example, to transfer biparietal-diameter and head/abdominal circumference measurements along with the related image; a gastroenterologist, urologist, or pulmonologist to convey a report of endoscopic findings, including annotation/labelling, measurement, and text descriptions of significant findings; or a pathologist to produce a histopathology report including measurements and coded descriptions linked to the DICOM-image coordinates of histopathologic findings. DICOM Supplement 23: Structured Reporting (as well as DICOM Supplement 15: Visible Light Image) specifies the SNOMED DICOM Microglossary as the controlled terminology for DICOM Code Sequence Data Elements.

The official site for SNOMED is www.snomed.org

The French translation project is described in the site www.crc-cuse.usherb.ca/kameleon/SNOMED/ *(no longer available in 2007: see now http://www.sogique.qc.ca/magazine/archives/vol14no2/entrevue.htm)*

## *UMLS*

In 1986, the National Library of Medicine, (NLM) began a long-term research and development project to build a Unified Medical Language System (UMLS®)."

The purpose of the UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources and to make it easy for users to link disparate information systems, including computer-based patient records,bibliographic databases, factual databases, and expert systems.

The UMLS project develops machine-readable "Knowledge Sources" that can be used by a wide variety of applications programs to overcome retrieval problems caused by differences in terminology and the scattering of relevant information across many databases.

The UMLS Metathesaurus is one of four knowledge sources developed and distributed by the National Library of Medicine as part of the Unified Medical Language System (UMLS) project. The Metathesaurus contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases and expert systems. **It preserves the names, meanings, hierarchical contexts, attributes, and inter-term relationships present in its source vocabularies; adds certain basic information to each concept; and establishes new relationships between terms from different source vocabularies.**

The Metathesaurus supplies information that computer programs can use to interpret user inquiries, interact with users to refine their questions, identify which databases contain information relevant to particular inquiries, and convert the users' terms into the vocabulary used by relevant information sources. The scope of the Metathesaurus is determined by the combined scope of its source vocabularies. The Metathesaurus is produced by automated processing of machine-readable versions of its source vocabularies, followed by human review and editing by subject experts.

The Metathesaurus is organized by concept or meaning. In essence, its purpose is to link alternative names and views of the same concept together and to identify useful relationships between different concepts.

- **Each concept or meaning in the Metathesaurus has a unique concept identifier** (CUI) which itself has no intrinsic meaning.
- **Each unique concept name or string in each language in the Metathesaurus has a unique string identifier** (SUI). Any variation in upperlower case is a separate string, with a separate SUI. The same string in different languages (e.g., English and Spanish) will have a different string identifier for each language.
- For English language entries in the Metathesaurus only, each string is linked to all of its lexical variants or minor variations by means of a common term identifier (LUI). (In the Metathesaurus, therefore, an English "term" is the group of all strings that are lexical variants of each other.)

All string and term identifiers are linked to at least one concept identifier. Different terms with the same meaning are linked to the same concept identifier. Thus in the Metathesaurus, strings are linked to terms and both strings and terms are linked to concepts.

| Concepts (CUIs) | Terms (LUIs) | Strings (SUIs) |
|---|---|---|
| C0004238 | L0004238 | S0016668 |
| (preferred)<br>Atrial Fibrillation<br>Atrial Fibrillations<br>Auricular Fibrillation | (preferred)<br>Atrial Fibrillation<br>Atrial Fibrillations | (preferred)<br>Atrial Fibrillation |
| | | S0016669<br>Atrial Fibrillations |
| | L0004327 | S0016899 |
| | (synonym)<br>Auricular Fibrillation<br>Auricular Fibrillations | (preferred)<br>Auricular Fibrillations |
| | | S0016900 |
| | | (plural variant)<br>Auricular Fibrillations |

For example, in the above table, the string "Atrial Fibrillation" and its plural "Atrial Fibrillations" have different string identifiers, but are linked to the same term identifier. "Auricular Fibrillation" and its plural "Auricular Fibrillations" are linked to a different term identifier. Since "Atrial Fibrillation" and "Auricular Fibrillation" have been judged to have the same meaning, their different term identifiers are linked to the same concept identifier.

The Metathesaurus also includes use information, including the names of selected databases in which the concept appears, and, for MeSH terms, information about the qualifiers that have been applied to the terms in MEDLINE. Information on the co-occurrence of concepts in MEDLINE and in some other information sources is also included.

The official site of UMLS is http://www.nlm.nih.gov/research/umls/umlsmain.htm

The use of UMLS is free for research purposes, provided you sign a Licence agreement.

The use of source vocabularies contained in the UMLS files is subject to specific restrictions. For example, Category 3 restrictions implies that

> *LICENSEE's right to use material from the source vocabulary is restricted to internal use at the LICENSEE's site(s) for research, product development, and s tatistical analysis only. Internal use includes use by employees, faculty, and students of a single institution at multiple sites. Notwithstanding the foregoing, use by students is limited to doing research under the direct supervision of faculty. Internal research, product development, and statistical analysis use expressly excludes: use of material from these copyrighted sources in routine patient data creation; incorporation of material from these copyrighted sources in any publicly accessible computer-based information system or public electronic bulletin board including the Internet; publishing or translating or creating derivative works from material from these copyrighted sources; selling, leasing, licensing, or otherwise making available material from these copyrighted works to any unauthorized party; and copying for any purpose except for back up or archival purposes.*

Category 3 applies to:

- Spanish, Portuguese, French, German translations of MeSH

- SNOMED
- ICD-10
- READ
- ACR when included. The UMLS coordinator wrote this message in May 1998:

> *The removal of the ACR codes from the UMLS Metathesaurus does not mean that there has been any change in their status or use here in the U.S. The codes that were present in previous editions of the Metathesaurus were not from the current edition. The American College of Radiology asked NLM to update to the current edition and told us that we should NOT distribute information from the previous edition. Due to other priorities, we were unable to add the new edition for 1998. We will probably do so in the future, but probably not for 1999.*

There are not specific restrictions for using:

- Transliterated Russian Translation of MeSH
- MeSH
- ICD-9-CM

## 4. Comparing different coding systems

The wide choice of different coding systems shows immediately that there is not a universal and optimum classification scheme. Each one has drawbacks.

An interesting study regarding this problem is by Campbell JR and Payne TH: "A comparison of four schemes for codification of problem lists", published in the Proc Annu Symp Comput Appl Med Care 1994:201-5. Here is the abstract:

> *We set out to evaluate the completeness of four major coding schemes in representation of the patient problem list: the Unified Medical Language System (UMLS, 4th edition), the Systematized Nomenclature of Medicine (SNOMED International), the Read coding system (version 2), and the International Classification of Diseases (9th Clinical Modification)(ICD-9-CM). We gathered 400 problems from patient records at primary care sites in Omaha and Seattle. Matching these against the best description found in each of the coding schemes, we asked five medical faculty reviewers to rate the matches on a five-point Likert scale assessing their satisfaction with the results. For the four schemes, we computed the following rates of dissatisfaction, satisfaction, and average scores: [table: see text]. From this analysis, we conclude that UMLS and SNOMED performed substantially better in capturing the clinical content of the problem lists than READ or ICD-9-CM. No scheme could be considered comprehensive. Depending on the goal of systems developers, UMLS and SNOMED may offer different, and complementary, advantages.*