



WORLD HEALTH ORGANIZATION

WHO/GPE/ICD/C/00.33

Distr.: LIMITED

ENGLISH ONLY

MEETING OF HEADS OF WHO COLLABORATING CENTRES
FOR THE CLASSIFICATION OF DISEASES

Rio de Janeiro, Brazil
15-21 October 2000

Electronic Maintenance of Clinical Classifications: Comparing Two Approaches

Stahl C¹, Walker SM¹, Garrett C², Truran D², Roberts R², Schopen M³

¹National Centre for Classification in Health, School of Public Health, Queensland University
of Technology, Victoria Park Road, Kelvin Grove, Queensland 4059

²National Centre for Classification in Health, Faculty of Health, University of Sydney,
Lidcombe

³German Institute for Medical Documentation and Information

This document is not issued to the general public, and all rights are reserved by the World Health Organization (WHO). The document may not be reviewed, abstracted, quoted, reproduced or translated, in part or in whole, without the prior written permission of WHO. No part of this document may be stored in a retrieval system or transmitted in any form or by any means - electronic, mechanical or other - without the prior written permission of WHO.

The views expressed in documents by named authors are solely the responsibility of those authors.

Electronic Maintenance of Clinical Classifications: Comparing Two Approaches

Stahl C¹, Walker SM¹, Garrett C², Truran D², Roberts R², Schopen M³

¹National Centre for Classification in Health, School of Public Health, Queensland University of Technology, Victoria Park Road, Kelvin Grove, Queensland 4059

²National Centre for Classification in Health, Faculty of Health, University of Sydney, Lidcombe

³German Institute for Medical Documentation and Information

This paper will give basic information about two different ways to store the International Classification of Diseases electronically. It aims to compare and contrast two approaches to the electronic maintenance of the ICD-10. The overall purpose in undertaking the comparison is to identify the strengths and weaknesses of each method in order to make recommendations to the World Health Organization regarding the maintenance and future development of classifications for which it has carriage.

The issues that will be addressed are as follows:

- Reasons for choosing the system
- Current uses and potential future uses
- Weaknesses and strengths.

Introduction

ICD-10 is the tenth revision of the International Statistical Classification of Diseases and Related Health Problems and is developed and maintained by the WHO. It is a statistical classification, which means that it contains a limited number of mutually exclusive code categories, which describe all disease concepts. The classification is hierarchical in structure with subdivisions to identify broad groups and specific entities. The main purpose of the International Classification of Diseases is to produce sets of internationally and nationally consistent information on morbidity and on causes of death, using established criteria and conventions. Up to the present time the WHO has not used an electronic method to maintain the ICD-10. Since 1996, WHO has issued annual updates to the ICD-10 in an attempt to keep the classification current. Managing the update process, rather than the production of an entirely new classification (as has occurred in the past) is a significant challenge. Such maintenance requires a system to keep track of changes made and to ensure consistency of approach in dealing with changes and the implementation of new codes. The WHO update process requires the submission of recommendations for changes to the classification from the international coding community to the WHO Update Reference Committee. All changes are documented using an Access database and sent to the WHO Collaborating Centres for the Classification of Diseases for consideration. At the annual Collaborating Centres meeting, the changes are ratified and subsequently the Evidence for Policy Unit in Geneva makes modifications to the classification. Changes are also documented on WHO's home page and the Collaborating Centres promulgate them to countries within their jurisdictions.

The National Centre for Classification in Health (NCCH) is the Australian Centre for excellence in health classification theory and an expert centre in coding systems. The centre is dedicated to supporting NCCH clients in their use of health classifications and related products and services. During 1998-1999, NCCH developed an Australian modification of the International Statistical Classification of Diseases and Related Health Problems (ICD-10-AM) for the classification of diagnosis and procedures. NCCH, through the

Australian's National Health Information Management Group (NHIMG), has agreed to the updating of this classification every two years to ensure the content remains current and the codes representative of Australian clinical practice. The Australian process for update incorporates public submissions from interested parties. These submissions are considered firstly by the NCCH, and subsequently by the Coding Standards Advisory Committee, which was established to assist the NCCH with issues relating to coding, the development of the classification and the Australian Coding Standards. Changes are approved by the National Health Information Management Group prior to implementation.

The German Institute for Medical Documentation and Information (DIMDI) is an institute within the scope of the German Federal Ministry of Health. DIMDI publishes official classifications (ICD-9, ICD-10, the procedure classification OPS-301), and also the German language editions of MeSH (Medical Subject Headings) and UMDNS (Universal Medical Device Nomenclature System).

Reasons for choosing the system

The electronic maintenance system designed for the Australian modification of the ICD-10 is a Visual Basic programmed Access database, which was developed by Essential Software on behalf of the NCCH, Sydney (NCCH, 2000a). Microsoft Access was chosen because it is part of the Microsoft Office Suite of programs, which was available on all computers at the NCCH and widely used in the health industry in Australia. Furthermore, it was anticipated that most potential customers would be familiar with Microsoft Office products. Access can easily be converted to SQL should the need arise. Combined with MS Word, the database can also support electronic and hardcopy publication. The main concerns raised with the choice of Microsoft Access, were whether it would allow multi-user access and whether it is robust enough to handle the volume of information required by the whole classification (over 100 Megabytes).

In contrast to the Australian ICD-10-AM Database, the maintenance of the German language edition of ICD-10 is through the use of an SGML document (Standard General Markup Language, ISO 8879) – a simple ASCII file with text and additional markup to indicate the logical structure of the classification. It was decided to use SGML for data storage and maintenance as it allows the automatic creation of different adaptations of the classification, different file formats (ASCII, RTF, HTML) or products like metadata files or update lists (Bryan, 1988). As ICD makes use of structures with rather fine granularity (nested lists, cross-referenced tables) and as access to these “granules” is needed both for typesetting and for structural transformations, SGML was chosen as it supports the fine granularity without additional implementation effort. SGML documents are open to structural changes which may become necessary during the long life time of the classification. Furthermore, they are independent of hardware or software, but standard processing software is available from different vendors. At the moment, the production system consists of an SGML-based editor (AuthorEditor), an SGML-based formatter (JADE), which supports style sheets written in the Document Style Semantics and Specification Language (DSSSL, ISO 10719), and an SGML-based transformation engine (BALISE), which uses a C-like language with an additional parser and interface for SGML documents. BALISE supports complex document composition or database-oriented applications.

Current uses and uses in the future

At this time the NCCH is the major user of the Australian database. Agreements have also been negotiated with software developers who develop coding products for the Australian market, and also with the US National Library of Medicine, which is assisting in a feasibility study to include the ICD-10-AM in the Universal Medical Language System (UMLS) (NCCH, 2000a).

Current uses of the Access based ICD-10-AM production system

- Production of the second and third editions of the classification (ICD-10-AM)
- Increased efficiency in search functionality and inbuilt data integrity with the development of business rules, reduction of human error and improved consistency
- Maintenance of ICD-10-AM

- Generation of addenda and errata
- Tracking changes to the Australian codes and Coding Standards over time
- Creation of mapping tables between editions of ICD-10-AM and between ICD-10, ICD-9, ICD-9-CM and ICD-10-AM
- Production of ICD-10-AM Browser
- Production of ASCII list for second edition ICD-10-AM
- Extraction of ICD-10-AM index terms from the database for inclusion in the UMLS metathesaurus and mapping to UMLS concepts.

Future uses of the Access based ICD-10-AM production system

- Publication of future editions of ICD-10-AM in both hard copy and electronic format including CD-ROM and ASCII list
- Creation of mapping tables between classifications and between editions of the same classification. Some functionality could be used to compare international versions of the ICD-10 as well
- Creation of subsets of the classification for specialty purposes eg Mental health and Allied Health
- Access to byproducts of the ICD-10-AM database on the World Wide Web
- Development of a standardized electronic format for the ICD-10
- Productions of ICD-10 database, eg the WHO version for the use of the WHO update Reference Committee
- Further development and refinement of an Australian Clinical Vocabulary
- Incorporation of the Australian Coding Standards in order to make this part of the classification searchable and consistent with the rest of the classification.

Current uses of the SGML-based ICD-10 production system

- Automatic transformation of ASCII texts of the German translation into an SGML document
- Further data entry for the first version of the German edition of ICD-10
- Production of the first German edition of ICD-10
- Maintenance of the classification
- Production of the second and third edition of the classification
- Production of errata and addenda for the second and third edition
- Extraction of terms for the German “Thesaurus of Diagnostic Terms”
- Extraction of German terms for inclusion into UMLS
- Creation of a defined subset of the classification for use in primary care (so-called ICD-10-SGBV)
- Automatic transformation into ASCII, RTF and HTML
- Automatic creation of the meta data file

Future uses of the SGML-based ICD-10 production system

- Publication of future editions in hard copy and electronic format
- Inclusion of the English and French versions to implement a multi-lingual system
- Sharing of consistency checks and production routines for these languages
- Enhancement to the production system by a document management system based on an object-oriented database

Weaknesses and strengths

Weaknesses of the Access based ICD-10-AM production system

The database is designed for use with at least a seventeen-inch monitor, otherwise the text of forms does not fit onto the screen as intended. The user can still work on the screen but will have scroll bars as is the standard with Microsoft products where the document is larger than the screen. The system requires a

minimum configuration of an i486 or Pentium processor-based personal computer (Pentium recommended) and 64 MB (128 MB recommended) of RAM. This is a fairly powerful system and may not be in general use in the coding community. Furthermore the system requires MS Windows 95, 98 or NT 4.0 and Access 97 and some third party ActiveX controls to operate. These controls are distributed royalty free in a runtime environment. Also required is at least 140 MB available hard disk space (NCCH, 2000b).

The searching functions in the Tabular List in the application front end of the database are not completely satisfying. It is possible to search for a code but not for words or descriptions. Certainly, there is the possibility to search in the data tables, or back end, of the database but then the user needs to have some knowledge about the tables and the variables in the database. Further development is planned to incorporate extra functionality in this area.

Strengths of the Access based ICD-10-AM production system

Microsoft Access is one of the most popular programs to handle databases worldwide. Furthermore the typical user does not need programming knowledge to use the database. Editing the ICD-10-AM classification with Microsoft Access is relatively simple and in a few days it is possible to be familiar with the most common and necessary functions.

The Australian ICD-10-AM database, combined with Microsoft Word, is a stand-alone system. A printed version of the classification, complete with typesetting conventions, is obtained by exporting a specified range of data to a Word template which will apply the appropriate formatting. If a change of formatting is required, the user would then modify the style settings in the Word template, thereby keeping formatting considerations separate from database requirements.

The browser and HTML documents are similarly generated. The browser export function will generate a Word document containing links between tabular, index and standards documents.

Selections of data are easily exported in a range of formats by means of in-built Access functions.

The structure of the ICD-10-AM database represents the original structure of the WHO-10 classification, which breaks each chapter into subchapters, blocks, categories and codes. Because of inconsistencies in the classification structure, predominantly in three chapters (Neoplasms, Musculoskeletal and External causes) which contain an additional level of code grouping, it was necessary to add this level to all chapters to maintain consistency. The key structure means that every entry in the database has a meaningful and unique identifier. The key contains all hierarchy information for an index term or a tabular code.

Basic rules to maintain the database integrity have been applied to the database. This ensures consistency in the maintenance of the classification and is considered one of the major advances compared with the previous manual update process. For example, it is impossible to add a 5th character code before the addition of a 4th character code. Similarly the index does not allow the addition of a 4th level indentation without the existence of a 3rd level indentation. For an example of the structure of the ICD-10-AM index, see Figures 1 and 2. Other database integrity assurance mechanisms include:

- Alphabetical order – All terms and modifiers are displayed in alphabetical order. For example,
A
Aarskog's syndrome Q87.1
Abandonment T74.0
Abasia (-astasia)(hysterical) F44.4
- When the attribute Not Elsewhere Classified (NEC) is chosen in the Index editing screen, the system will automatically allocate NEC as the value and place it in the correct sequence following the main term. For example,
Autoimmune disease (systemic) NEC M35.9
- When a non-essential modifier is added in the Index editing screen, the system will automatically allocate parentheses. For example, Macrotonia (congenital)(external ear) Q17.1

The Australian ICD-10-AM database allows authorized users at the NCCH to make required updates to the classification, to export the index and the tabular list to Microsoft Word and to develop report and query functionality within Access. The database has been implemented in a multi-user environment. All edits made in the database are written to a log file which records data before and after a change, user name, date, reason for the edit and any other information the user considers relevant. This file can then be edited to form an errata or addenda, which was previously an onerous task (NCCH, 2000b).

It is not anticipated that the database will be used by coders themselves, rather only by those responsible for the development and maintenance of the classification. This is because maintenance of version control and code standardization is necessary, so that local codes and coding standards do not proliferate. Coders will instead use the ICD-10-AM browser product as well as the hard copy publications.

Extraction of terms for inclusion in the UMLS metathesaurus is an important feature. Previously, the richness of medical terms contained in the ICD Index has not been readily accessible – the index was alphabetical and searching for ‘related’ terms was time-consuming and laborious. The database, being relational, ‘links’ related terms and allows them to be quickly and reliably located. For instance, the ICD-10-AM has approximately 20 000 codes, but in excess of 80 000 ‘terms’ in the Index and the essential and non-essential modifiers. This means that there is considerably more clinical information available in the classification than merely accessing the codes and rubrics permits.

Weaknesses of the SGML-based ICD-10 production system

Experience in the application of SGML is not as widespread as experience in the application of relational database technology. While the increasing use of the WWW has considerably improved this situation, we still expect the expertise in this area to grow through the growing use of XML, as both HTML and XML are applications of SGML.

SGML-based software is not part of standard office suites.

At the moment the data entry part of the production system allows only single user input. As ICD-10 is split into chapters, this is not a major obstacle as another user could work on a different chapter. Nevertheless, we hope to overcome this weakness by the integration of a document management system into the production environment. This would allow multi-user input based on record locking. Most of these document management systems are based on object oriented databases so that the hierarchical structure of the classification can be easily maintained even beyond the subcategory level (Schopen, 1998).

At the moment the Tabular List of ICD-10 and the Alphabetical Index are only linked during the run of a special consistency check program and during the run of a special printing program. However, they are not linked during data entry. This is also to be overcome with the integration into a document management system.

Strengths of the SGML-based ICD-10 production system

Although software conformant to the SGML standard is not part of usual office suites, a complete production system (editor, formatter, transformation engine) can be set up using public domain software available free of charge via the Internet (Schopen, 1998).

The structure of the ICD can be represented effectively beyond the subcategory level. Structures like nested lists, tables or cross-references to subclassifications are tagged in detail. They are not only accessible for editing but also for effective searches, as text searches can be restricted to certain elements or certain element hierarchies. Even more complex queries become possible with the transformation engine, which can keep the document tree of the ICD in memory and allow arbitrary navigation along this tree (Schopen, 1999). For an example of the German index structure, see Figure 3.

As SGML documents must be validated against a document type definition (DTD), a complex set of rules can be defined for the structure of the document. During data entry these rules are enforced. Eg. it is only possible to insert subcategories within categories, a level 3 indentation in the alphabetical index only within a level 2 indentation, or an asterisk code only following a dagger code. Consistency checks which cannot be formulated in the DTD can be programmed with the transformation engine. Eg. it is checked that each exclusion note or each entry in the alphabetical index references an existing ICD code or that *see* or *see also* references in the alphabetical index refer to existing entries with correct wording.

Together with the DSSSL standard, the production of hard copies is supported in really high quality. As layout should never be part of an SGML document but should always be added later by style sheets based on the element structure, consistent formatting is inherent to the SGML approach.

Sorting is supported in a very flexible way by the use of so-called SGML entities, which support different replacement texts for production and for sorting. Thus it is possible to sort diacritics, Greek characters or Roman numbers without any problems or data duplication. Certain words like “with” as the starter of an entry can easily be kept out of the sorting process in order to list them in the index before all other entries. Furthermore, entities can be used to keep standard texts consistent throughout the entire classification.

A document with a fine granularity is open to structural changes or to structural transformations. At the moment the German ICD-10 is offered in three versions: book version (as printed), computer version (with full titles, making explicit all codes to be built by subclassifications), and ICD meta data (codes and titles, blocks and chapters for relational databases with edits for patient gender and additional data fields for statistical tabulation). These versions are created automatically from a single SGML document, thus avoiding data duplication and manual work. Furthermore, a subset of ICD-10 for morbidity coding in primary care is extracted automatically from this SGML document and also offered in these three versions. Thus, consistency of this primary care version with the full ICD is ensured. Another kind of transformation is the generation of files with update and corrigenda information. They can be extracted automatically from the SGML documents and be formatted by a style sheet in a way that insertions and deletions can be controlled by the user (Schopen, 1999).

SGML tags can be regarded as containers for information. As the language of this information does not matter and as the structure of ICD-10 is identical in all languages, it should be possible to maintain the ICD in several languages using this technology and to share production programs for these languages. First tests with the English and French versions of ICD-10 were successful.

Figure 1: Example from the ICD-10-AM Alphabetical Index in traditional book-like view

<p>Abnormal, abnormality, abnormalities – see also Anomaly</p> <ul style="list-style-type: none"> - alphafetoprotein R77.2 - amnion, amniotic fluid O41.9 - - affecting fetus or newborn P02.9 <p>Perforation, perforated (nontraumatic)</p> <ul style="list-style-type: none"> - bowel K63.1 - - fetus or newborn P78.0 - colon K63.1
--

Figure 2: Example from the Australian ICD-10-AM database Alphabetic Index data structure

Table: Index_Term

Key	Level
MAIN:Abnormal, abnormality, abnormalities :10Abnormal, abnormality, abnormalities	0
MAIN:Abnormal, abnormality, abnormalities :10Abnormal, abnormality, abnormalities :70alphafetoprotein	1
MAIN:Abnormal, abnormality, abnormalities :10Abnormal, abnormality, abnormalities :70amnion, amniotic fluid	1
MAIN:Abnormal, abnormality, abnormalities :10Abnormal, abnormality, abnormalities :70amnion, amniotic fluid :70affecting fetus or newborn	2
MAIN:Perforation, perforated :10Perforation, perforated	0
MAIN:Perforation, perforated :10Perforation, perforated :70bowel	1
MAIN:Perforation, perforated :10Perforation, perforated :70bowel :70fetus or newborn	2
MAIN:Perforation, perforated :10Perforation, perforated :70colon	1

Table Index_Term_Note

Note Type	Note Text
See Also	Anomaly
Code	R77.2
Code	O41.9
Code	P02.9
Non Essential Modifier	nontraumatic
Code	K63.1
Code	P78.0
Code	K63.1

Figure 3: Example from the SGML-based Alphabetical Index data structure

```
<Index>
  <L0 ID="000100"><Entry><Text>Abnormal, abnormality, abnormalities</Text>
    <SeeAlso TO="000303">Anomaly</SeeAlso></Entry>
    <L1 ID="000101"><Entry><Text>alphafetoprotein</Text><Code>R77.2</Code></Entry>
    </L1>
    <L1 ID="000102"><Entry><Text>amnion, amniotic fluid</Text><Code>O41.9</Code></Entry>
      <L2 ID="000103"><Entry><Text>affecting fetus or newborn</Text>
        <Code>P02.9</Code></Entry>
      </L2>
    </L1>
  </L0>
  <L0 ID="020243"><Entry><Text>Perforation, perforated (non traumatic)</Text></Entry>
    <L1 ID="020244"><Entry><Text>bowel</Text><Code>K63.1</Code></Entry>
      <L2 ID="020245"><Entry><Text> fetus or newborn</Text><Code>P78.0</Code></Entry>
      </L2>
    </L1>
    <L1 ID="020256"><Entry><Text>colon</Text><Code>K63.1</Code></Entry>
    </L1>
  </L0>
</Index>
```

Conclusion

Both ways to store the International Classification electronically add considerably to our knowledge of the ICD-10 classification and its structure. The discipline of translating the classification from hard copy to an SGML document or a database exposes the complex relationships within this classification which has evolved over many decades and withstood the test of international use over time. Only through these electronic refinements will we be able to manage efficiently the updating of the classification to maintain its currency and applicability for morbidity and mortality reporting purposes.

An electronic version of the classification is vital as a foundation for the electronic health record. Without the ability to translate words into code for storage, transmission and analysis, we cannot hope to move ahead in improving the collection and use of health data. The means of electronic maintenance of the classification reviewed in this paper are essential ingredients for the implementation of policy on electronic health records and translation of health data into meaningful information. This will allow us to bridge the gap between languages, not only German and English, but also the understanding of health terms used by administrators, clinicians and consumers.

References

1. Bryan, M. (1988) *SGML: An author's guide to the Standardized Generalized Markup Language*. Addison-Wesley Publishing Company: Wokingham.
2. NCCH (1998) *The ICD-9-CM/ICD-10-AM Mapping Tables*, Internal Report. NCCH: Sydney.
3. NCCH (2000a) *Development of ICD-10-AM Database: Report to National Priorities and Quality Branch, Commonwealth Department of Health and Aged Care*. NCCH: Sydney.
4. NCCH (2000b) *ICD-10-AM User Guide version 1.0.0*. NCCCH: Sydney.

5. Schopen, M. (1998) *Electronic Publishing of the German Language Edition of ICD-10, DIMDI*. In: Meeting of Heads of WHO Collaborating Centres for the Classification of Diseases, Paris, 13-19 October 1998. Geneva, World Health Organization, 1998 (unpublished document WHO/GPE/ICD/C/98.24; available on request from Global Programme on Evidence for Health Policy, World Health Organization, 1211 Geneva 27, Switzerland)

6. Schopen, M. (1999) *Electronic Publishing of the Alphabetic Index to ICD-10*. In: Meeting of Heads of WHO Collaborating Centres for the Classification of Diseases, Cardiff, 17-22 October 1999. Geneva, World Health Organization, 1999 (unpublished document WHO/GPE/ICD/C/99.33; available on request from Global Programme on Evidence for Health Policy, World Health Organization, 1211 Geneva 27, Switzerland)