



WORLD HEALTH ORGANIZATION

WHO/GPE/CAS/C/01.65

Distr.: LIMITED
ENGLISH ONLY

MEETING OF HEADS OF WHO COLLABORATING CENTRES
FOR THE FAMILY OF INTERNATIONAL CLASSIFICATIONS

Bethesda, MD, USA
21-27 October 2001

**DEVELOPING A DATABASE VERSION OF THE INTERNATIONAL
STATISTICAL CLASSIFICATION OF DISEASES AND RELATED
HEALTH PROBLEMS, TENTH REVISION, CLINICAL MODIFICATION
(ICD-10-CM)**

WHO Collaborating Center for the Family of International Classifications for North America

David J. Berglund, M.D., M.P.H.
Medical Officer, National Center for Health Statistics
Centers for Disease Control and Prevention
Hyattsville, Maryland, USA

Lizabeth J. Fisher, R.H.I.A.
Medical Systems Specialist
National Center for Health Statistics
Centers for Disease Control and Prevention
Hyattsville, Maryland, USA

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.

Abstract:

A database version of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Clinical Modification (ICD-10-CM), is being developed. The purpose of this activity is to support various users who have an interest in this electronic format in the United States, including many in the health care industry, as well as the academic community, and representatives from other government agencies.

This report will review potential uses and users of a database version of ICD-10-CM, and benefits for electronic data interchange, as well as mapping with clinical terminologies. An enterprise Relational Database Management System (RDBMS) is used, SQL Server 2000. For browsing and viewing, a client computer uses Word 2000 with additional specially developed software.

This paper will give a comparison of the RDBMS for maintaining ICD-10-CM with the proprietary Folio Infobase system used to maintain ICD-9-CM. The database system will enable output to multiple different formats needed by users. For the long term, the system must allow easy editing and maintenance of the codes. This is still in development. Use of open formats, and capability for integrity checking, are significant advantages.

1. Introduction and background

At the same time as the National Center for Health Statistics (NCHS) has been working on development of a clinical modification of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10-CM), it has been recognized that a maintenance system for ICD-10-CM will be needed. The system currently used for maintaining ICD-9-CM is Folio Views, which has a proprietary Infobase format. It has had some benefits, of enabling use of hyperlinks, and structured searches, while handling free text mixed with structured data well. However, it also has significant weaknesses, in that the format is not open, and only word processor output is supported. Yearly updates have required significant manual work to generate addenda. The NCHS Division of Information Technology (DIT) has not supported use of this program. Finally, the Folio program has been discontinued, so it is not clear how well supported it will be in the future.

Many non-technical users have been happy with the Folio interface. However, technical users who want to make use of the ICD in databases have been very unsatisfied. Thus, it appears to be most appropriate to provide ICD-10-CM in Folio format for those who find it useful, but also provide it in other formats to enable ready use by others. It is relatively straightforward to move data into Folio format from either SGML or database formats, so one of these may then be used for maintenance. There is also a need to support markup language formats, based on a legislative requirement.

Development of ICD-10-CM so far has primarily used WordPerfect, making use of document styles to specify structure. Data conversion to other formats will be needed. The original creation of the Folio Infobase files for ICD-9-CM had taken the word processor files, and used the styles to determine structure while converting to a Folio Infobase. A similar automated data conversion will need to be used, for moving the data into the format chosen.

2. Potential systems considered

In considering options for a maintenance system for ICD-10-CM, it was helpful to consider approaches to electronic maintenance of clinical classifications used elsewhere. The German ICD-10 is maintained in the SGML format using Author Editor, and other tools (JADE and BALISE). The Australian ICD-10-AM is maintained using a Microsoft Access database, and custom tools. An excellent paper from the 2000 Centre Heads meeting in Rio de Janeiro (WHO/GPE/ICD/C/00.33) compared these two approaches. Another option ultimately chosen was an enterprise Relational Database Management System (RDBMS).

SGML is widely used in the publishing industry, and certain editors for it are very powerful. WordPerfect has an SGML mode, but it has very limited capabilities. For example, it does not support having a Document Type Definition in multiple files, nor does it support the use of

SGML markup for insert and delete to enable tracking document changes. Furthermore, it uses a nonstandard compiled format for style sheets and other document information. The limitations prevent it from being a viable option. More powerful SGML editors such as FrameMaker and Author Editor have not been supported at NCHS, and would require significant training and support to implement. SGML appears to require significant technical expertise to take advantage of its capabilities.

Microsoft Access is now in use at NCHS and is supported. Some potential concerns raised about use of Access by the Australian Center, at the Rio meeting (WHO/GPE/ICD/C/00.33), were support for multiuser access, and whether it would be robust enough for the volume of information. For ICD-10-CM, there has been addition of a sixth digit, and significant amounts of content, which would increase the level of concern about whether it would be robust enough. Problems could potentially manifest as simple system slowing, or with overt program failures. Beyond this, to meet NCHS' needs, significant custom programming would be necessary. It is not clear whether such programs created for the current version of Access will work properly with the next version.

An enterprise Relational Database Management System (RDBMS) was another option considered. These allow use of Structured Query Language (SQL), and generally provide more capabilities for multiuser support, and are more robust. An example is the Oracle database, which is used with Apelon tools to maintain the National Library of Medicine (NLM) Unified Medical Language System (UMLS) Metathesaurus. The NCHS DIT supports two enterprise RDBMS programs, Sybase and Microsoft SQL Server.

3. Planned uses for system, and criteria for choosing system

In order to help review the options available and determine capability to meet needs, specific criteria were stated. The NCHS DIT was also consulted. The system needed to meet the following abilities.

- a. Enable browsing and searching; use of hyperlinks; interface to be similar to Folio
- b. Enable editing for maintenance; track all changes
- c. Enable automated creation of addenda when updating
- d. Enable output to multiple formats (ASCII text, database, markup language, as well as formats for print such as Word and PDF)
- e. Support by DIT (ideally)
- f. Potential to allow access by users external to NCHS for browse and search; e.g., support web based browsing, or access to a database

4. System chosen: enterprise Relational Database Management System (RDBMS)

The system ultimately chosen based on review and input from NCHS DIT was Microsoft SQL Server 2000, with the operating system supported by NCHS being Windows NT 4, with Internet Information Server. As part of the system, Microsoft Word will be used for viewing and editing, via a custom application (written using COM+ and Visual Basic for Applications). The data will be converted to XML fragments for viewing in Word. To speed response time the XML fragments can be pre-created, and placed on the server for direct access. This should work well because the system will probably be updated only once or twice a year. Updates will take somewhat longer, but viewing and browsing will be faster.

Data conversion was automated. Data in WordPerfect files was converted to Word 2000, exported to HTML/XML, and then parsed and placed into database tables in a staging area. Following further review and manipulation, the data was moved into the final database.

5. Weaknesses and strengths of system

The use of SQL Server for maintenance, along with Word and additional custom tools, has both weaknesses and strengths when compared with other options and with the Folio Views program.

a. Weaknesses

- i. Data located on the server takes some time to be transferred over the network, so the entire file is not instantly accessible (as was the case for Folio). Also, data fragments do not flow smoothly as the Folio Infobase did; thus, multiple categories may not be visible at the same time.
- ii. Handling special characters in the database was problematic (ultimately it was decided to use limited markup language to handle these). Similar problems would have been encountered with Folio, as it does not support Unicode, and potentially would have had difficulty with Greek characters.
- iii. Development has been done by an outside contractor; it has taken longer than initially projected.

b. Strengths

- i. Output to multiple open formats will be straightforward, with much better interoperability than Folio supplied.
- ii. Searches directly make use of the database (rather than the XML fragments). SQL Server has sophisticated search capabilities, even better than those of Folio.
- iii. Data integrity checking can be supported, which was not feasible in Folio due to the proprietary data format. Some checking can be done during the data conversion, which depends on the WordPerfect styles being correct; this will potentially allow errors to be caught and corrected.
- iv. The transaction support of the enterprise RDBMS can fully track changes

- v. to the database, for maintenance purposes.
- v. Data being on the server allows DIT to back it up regularly without users intervention necessary. SQL Server is supported by the NCHS DIT.

6. Current status and future plans

The database system has been developed to an early beta status. Many early issues have been addressed, but others still need to be resolved before it will be useable. Parts of the system that work reasonably well include the data conversion routines, and the search and browse capabilities. Components working in more preliminary form include editing via Word with submission to the database, and output to other formats. Development of the database system is currently on hold, while ICD-10-CM content is finalized. It is anticipated that work will resume in the next fiscal year, which begins October 1, 2001.

The NLM plans to include ICD-10-CM in the UMLS Metathesaurus. This will involve moving the data into the Oracle database that is used for maintaining UMLS, along with additional integrity checks. Full error checking for data integrity is needed (work so far has included some error checking, but more is needed). Mapping to other systems such as the Metathesaurus, and vocabularies such as SNOMED, will be much easier to accomplish with the database version of the ICD-10-CM.

It is anticipated that yearly updates to ICD-10-CM will be performed, using the same process for obtaining public input as for ICD-9-CM. Expected advantages of the new system for updating include the automated creation of the human readable addenda, along with creation of a file including the changes, which could be used for updating database systems that incorporate ICD-10-CM.

7. Conclusion

NCHS will be moving to an enterprise RDBMS for maintaining ICD-10-CM. This has both advantages and disadvantages. It will be more interoperable with other systems than Folio has been, support output to multiple open formats, and enable integrity checking. However, the transition will take some time. The system chosen has both strengths and weaknesses, but it should fully meet the needs for both internal and external use of ICD-10-CM.

References

Stahl C, Walker SM, Garret C, Truran D, Roberts R, Schopen M. Electronic Maintenance of Clinical Classifications: Comparing Two Approaches. Meeting of Heads of WHO Collaborating Centres for the Classification of Diseases, Rio de Janeiro, Brazil, 15-21 October 2000. (WHO/GPE/ICD/C/00.33)

National Center for Health Statistics. Report of ICD-10-CM Future Publishing System Alternatives. Internal Report. June 2000.