European Statistics on Causes of Death – COD
Methodological information

IRIS: a language independent coding system for mortality data
Lars Age JOHANSSON (National Board of Health and Welfare - Sweden)
Gérard PAVILLON (National Institute for Health and Medical Research Inserm-CépiDc-France)

Introduction
Mortality data are widely used to describe and compare health characteristics between countries. The International Classification of diseases (ICD), published and updated by the World Health Organisation (WHO), is used by most countries to code causes of death [OMS92]. ICD is now more than one century old and the last revision of this classification (ICD10), reflecting the increasing knowledge of medicine, is so complex that only classification experts can use it properly.

Mortality statistics are mainly based on the concept of underlying cause of death, defined by WHO as the "a) the disease or injury which initiated the train of morbid events leading directly to death, or b) the circumstances of the accident or violence which produced the fatal injury". The selection of the underlying cause of death is based on the diagnoses reported by the certifier, usually a physician. The selection is directed by a intricate set of rules and guidelines, also included in the ICD. Some instructions are very detailed and only apply to a specific condition, but others are expressed in general terms and are open to various interpretations. To ensure consistent application of the ICD instructions, producers of mortality statistics employ professional medical coders (nosologists). Due to the complexity of the rules, and because the rules presuppose an extensive knowledge of possible causes and complications of the conditions mentioned on the death certificate, training a nosologist takes many years.
In this context, quality and comparability of data are essential [Eurostat01b]. Comparisons between medical coders show important variations due to interpretation differences or errors. By using a common software, countries would decrease the number of differences due to various interpretations of the ICD rules and guidelines.

Software for coding and selection have now been available for a long time. The US National Centre for Health Statistics introduced automated selection of the underlying cause in the late 1960ies. Further modules and functions have been added, and now the NCHS system, called the Mortality Medical Data System (MMDS), is used by many other countries [ICE99] [ICE01] [ICE06]. The MMDS is divided into three functional parts: coding of medical causes of death reported by the physician (MICAR), selection of the underlying cause of death (ACME) and compilation of multiple causes of death (TRANSAX). ACME is now considered as the de facto standard for countries using automated coding systems. Briefly, ACME selects and modifies the ICD code of the underlying cause of death from the set of ICD codes representing the conditions mentioned on the death certificate. This function is independent of the language used to report causes of deaths. MICAR module prepares ACME input by assigning ICD codes to the medical terms reported by the physician. This function is highly language dependent. Even if MICAR performs perfectly in English, many problems appear when it is applied to other languages. Consequently, while the MMDS has been fully applied by English-speaking countries (US, UK, Australia), only one part of it, namely ACME, is used by non-English-speaking countries (Sweden, Brazil, France, ...). Using ACME without MICAR has its problems, however, since ACME does not work properly unless the coding of the medical expressions follows a set of very specific instructions. MICAR applies these instructions automatically, but countries that do not use MICAR have to apply them manually, or to include the functions of MICAR in their own coding software. It is generally agreed that a common automated coding system, with modules for both text coding and selection of the underlying cause, would be the best way to improve mortality data quality and comparability. However, the problem of language is still a major obstacle.

**Objective**

The problem of MICAR with non-English languages is closely linked to the English dictionary of diagnostic expressions (medical terms). Each diagnostic expression is linked to an Entity Reference Number (ERN). ERN have been used instead of ICD codes in order to keep as much information as possible: several ERN may correspond to one ICD code. The expressions and the ERN, however, cannot always be mapped to similar non-English expressions. There might also be expressions in the non-English language that do not
correspond to any expression in the English dictionary. This all corresponds to the well-known difficulties of translation from one language to another.

In order to cope with this problem, we decided to study the possibility to implement the coding function (MICAR) in a language-independent way. The first objective is to produce a program that can be used with different languages by just changing the dictionary of causes of death. The second aim is to stay fully compatible with MMDS.

**Method**

At first, we planned to rewrite MICAR, and when doing so to separate the language and the algorithmic aspects [Eurostat01a]. It would then be easier to apply the program to different languages. This task proved to be very complex and therefore the compatibility between MICAR and the language independent program could not be guaranteed. Also, it would be difficult to synchronize the future development of these two different pieces of software.

The present project is based on a different idea. We assume that most ICD codes can be mapped to a single ERN and that this mapping has no visible effect on the selection of the underlying causes of death. ICD codes are international and they are defined in different languages. Therefore it seems possible to enter MICAR via ICD codes provided that a module processes the mapping from each ICD code to ERN code. With this approach the language aspects are present only in the dictionary, which is, of course, different for each language. This dictionary maps diagnostic expressions to ICD codes, and the system then translates ICD codes to ERN. MICAR, embedded as a component in the system itself, is called to apply the instructions for ACME input coding, and eventually produce a set of ICD codes for ACME. Finally, the system calls ACME which selects the underlying cause.

This system, called Iris, reaches the two objectives:

- It is language independent because linguistic aspects are concentrated in the dictionary using international definition of ICD codes;
- It is MMDS compatible since it includes MICAR and ACME modules as components. Each modification of these modules is immediately transferred in Iris.
We have also designed and developed specific features in order to help to adapt and use Iris.

**Language standardisation.** An integrated tool in Iris allows the pre-processing of language expressions in order to reduce the size of the dictionary. National experiences show that there is numerous ways to express a disease corresponding to a single ICD code. For instance the code "I21.9" corresponding to "Acute Myocardial infarction" is expressed by 483 different expressions in French. It is essential to be able to simplify automatically the original expressions so that the dictionary includes only standard expressions. The dictionary is then easier to maintain and to keep consistent. Iris offers a standardisation tool driven by Regular Expressions that the user can parameterize according to its needs.

**National adaptations.** Original Iris displays screens written in English. The program has been written so that specific national adaptation can be made. Messages and screens are set apart from the program logic so that the former can be translated without altering the latter.

**Interactivity.** Iris has been designed to cover the whole process from coding to underlying cause selection in one step: the coder can enter causes of deaths and immediately get the underlying cause of death. Of course, Iris also allows batch processing for large files of death records.

**Explanations.** On request, Iris can provide explanations on each step of the process. This aspect is essential when processing is complex to explain each step to the coder. It also allows using Iris as a pedagogical tool to teach causes of death coding. Finally, it makes available the expertise included in the software.

**Open software policy.** Given the objectives of the Iris project, it was important to keep Iris an international product. Iris evolutions are decided by a user group opened to each country using Iris. Iris has been referenced by the Agence pour la Protection des Programmes (http://app.legalis.net) with the following InterDeposit Digital Number (IDDN): FR.001.170001.000.S.P.2007.000.31235.
Results

Several tests have been made on Iris.
- Between MMDS and Iris by Hungary and Sweden;
- Between Iris and manual coding by Spain;
- on language standardisation by Germany.

The English international version of Iris is in the final step and will be available by May 2008. A German version will be implemented in 2008 in several Länder. A Swedish and a French version will be developed in 2008.
References


[ICE06] Proceedings of the International Collaborative Effort on Automating Mortality Statistics, Centers for Disease Control and Prevention, National Center for Health Statistics, November 2006