Chapter 6: Processing of data

J. Ferlay

There were 313 cancer registries that replied to the invitation to participate by submitting data for Volume IX of *Cancer Incidence in Five Continents* (CI5). Each cancer registry sent its own data, except the registry members of the Surveillance Epidemiology and End Results (SEER) and of the National Program of Cancer Registries (NPCR) of the United States, and six Canadian registries that provided their data in a single data submission.

The data were generally sent to the IARC secretariat through electronic mail or on CD-ROM for large files, and as usual, were accepted in any electronic format (text files, Excel™ spreadsheet, database files, etc.) and with any file layout. Contributing registries were also invited to send all their data containing all malignant and non-malignant diagnoses collected. In addition, official cancer mortality data for the reference period, and population data, ideally for each calendar year of the reference period, should also be provided and checked by the IARC secretariat. This resulted in the manipulation of more than thirty million individual records, and in the production of 389 preliminary datasets (including different ethnic groups) to be examined carefully by the editors. A procedure for data validation and storage, summarised in Figure 6.1, was established to handle the large amount of data submitted for the project.

Incidence data

The incidence data were submitted as listings of individual anonymous cases with at least the following variables:

1. A registration number that identifies the patient or the case
2. Sex
3. Ethnic group or race (optional)
4. Age and/or birth date
5. Date of incidence
6. Site of the tumour (topography)
7. Morphology of the tumour
8. Behaviour of the tumour
9. Most valid basis of diagnosis

A description of all the codes used for the variables had to be provided with the data. However, it was not unusual for the code values to not match the description provided or for the coding information to be missing. In that case, the registry was asked for clarification and to provide the correct codes if necessary. This is particularly important when computing the percentage of microscopically verified (MV) or death certificate only (DCO) cases used for the editorial process: a misinterpretation of the basis of diagnosis code could lead to a false picture of the data quality.

Conversion into ICD-O-3

The first stage of the editorial process is to examine the incidence data using the standard IARC-CHECK program. This requires the data to be coded into a full (Topography and Morphology) ICD-O-3 coding schema. A large majority (230) of the registries submitted data according to this classification, but the others (72) had to be converted before they could be handled by the program. Several datasets included both ICD-O-1 and ICD-O-2, not to mention ICD-9 and ICD-10 codes (for coding of the DCO cases) depending on the year of incidence of the case, so that they had to be split into two or more files, each piece of the puzzle being converted using the appropriate program (Figure 6.2). These preliminary conversions proved to be particularly valuable in detecting incompatibilities between ICD-9 or ICD-10 codes and ICD-O morphology and behaviour (Table 6.1), which were transmitted back to the cancer registry for review and correction.

Figure 6.1. Processing of cancer registry data to generate *Cancer Incidence in Five Continents Volume IX*

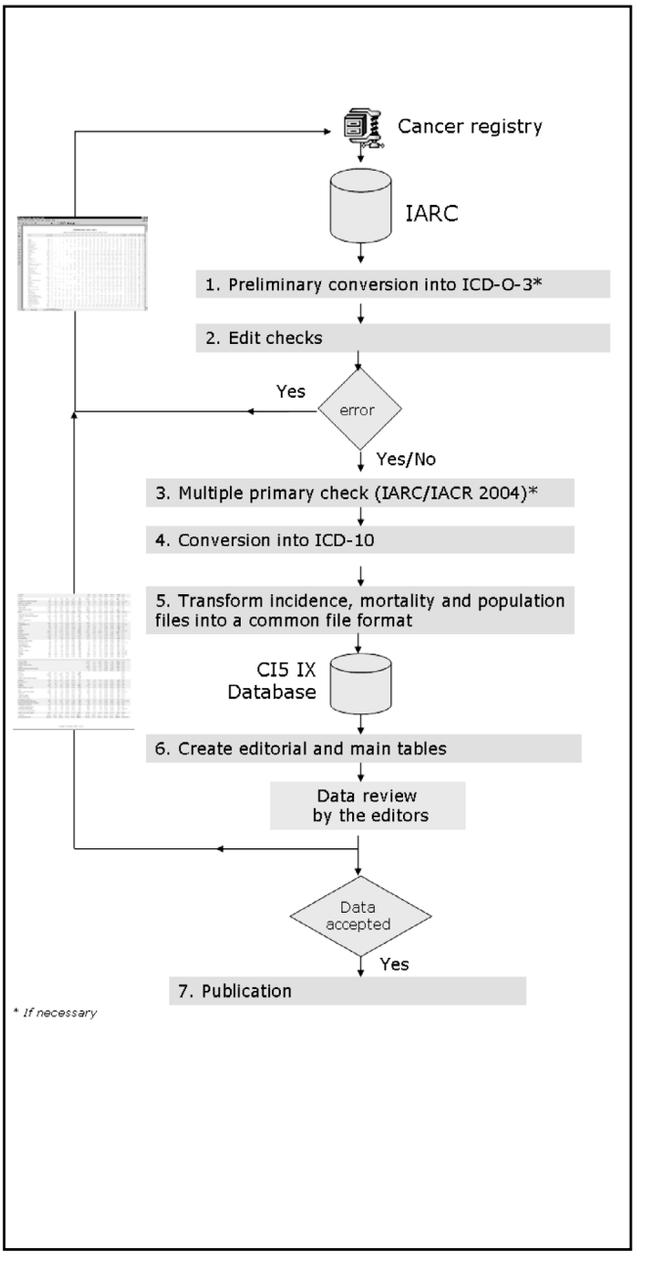

Table 6.1 Examples of unlikely ICD-10 site/ICD-O-2 morphology combinations

ICD-10		ICD-O second edition (correct code if incorrect)
C22.0	Hepatocellular carcinoma	8010/3 Carcinoma, NOS (8170/3)
C45.0	Mesothelioma of pleura	8050/3 Papillary carcinoma, NOS (9050/3)
C46.0	Kaposi sarcoma of skin	8140/3 Adenocarcinoma, NOS (9140/3)
C91.0	Acute Lymphoid Leukaemia	9823/3 Chronic Lymphocytic Leukaemia (9821/3)
C81.9	Hodgkin disease, NOS	9670/3 Malignant lymphoma, small lymphocytic (9650/3)
C43.9	Melanoma of skin, NOS	8090/3 Basal cell carcinoma (8720/3)
C34.9	Lung (primary cancer)	8140/6 Adenocarcinoma, metastatic (8140/3)
C53.9	Uterine cervix, malignant	8070/2 Squamous cell carcinoma in situ (8070/3)

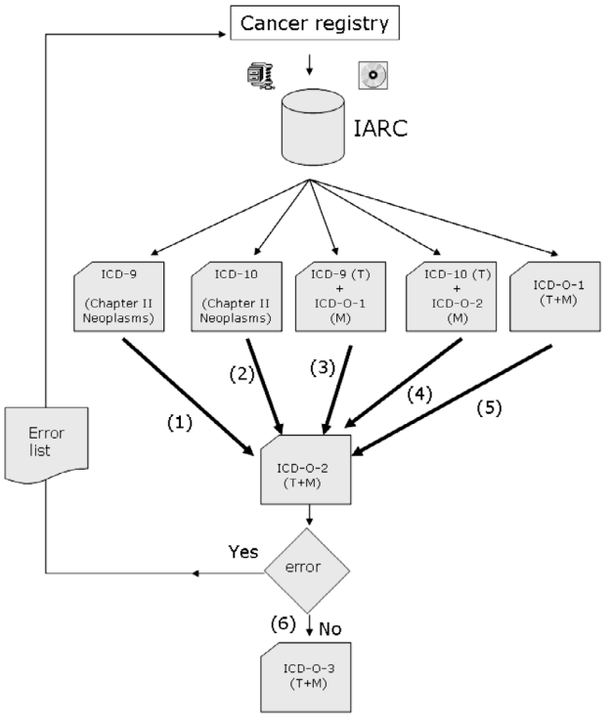

Figure 6.2 Preliminary conversions of cancer registry data
(1) ICD-9 (1975), chapter II neoplasms, to the second edition of ICD-O (1990) conversion program.
(2) ICD-10 (1990), chapter II neoplasms, to the second edition of ICD-O (1990) conversion program.
(3) ICD-9 (1975), chapter II neoplasms for topography and the first edition of ICD-O (1976) for morphology to the second edition of ICD-O (1990) conversion program.
(4) ICD-10 (1992), chapter II neoplasms for topography and the second edition of ICD-O (1990) for morphology to the second edition of ICD-O (1990) conversion program.
(5) First edition of ICD-O (1976) to the second edition of ICD-O (1990) conversion program.
(6) Second edition of ICD-O (1990) to the third edition of ICD-O (2000) conversion program.

For Finland, the site was coded to ICD-7 (1955) and the morphology using a modified version of the Manual of Tumor Nomenclature and Coding by the American Cancer Society (1951). A specific conversion program had been developed within the framework of the NORDCAN project, a collaboration between the IARC and the Nordic Association of Cancer Registries.

Although the ICD-O-3 gives clear instructions that behaviour codes /6 and /9 should not be used by cancer registries (page 27 of ICD-O-3), these codes appeared in few datasets, giving rise to problems with respect to the corresponding topography code. Usually, this was assumed to represent the site of the primary tumour. Where this was evidently not the case (carcinomas in lymph nodes, in bone, etc.), a listing of such cases was sent back to the registries with a request for clarification. As a last resort, they were recoded to topography C80.9 (primary site unknown).

Checking
Once a dataset had been converted into ICD-O-3, or if it had been originally coded using ICD-O-3 codes, it was submitted to the IARC-CHECK program, which performed the following edits:

1. Code verification

- Sex
- Incidence date and, if provided, birth date
- ICD-O-3 topography and morphology

2. Consistency between items

- Age versus birth/incidence dates
- Sex versus site
- Sex versus histology
- Age versus site
- Age versus histology
- Site versus histology
- Basis of diagnosis versus histology

Registries submitting data for Vol. IX were invited to run their own data through the IARC-CHECK program prior to submission, and a number of contributors did so (143), particularly the users of the CanReg-4 software (30), which includes the same edits. However, these datasets were automatically re-checked by the IARC secretariat. All errors or unlikely or rare combinations of items were sent back to the cancer registry for verification. The corrections or new submissions were then incorporated, converted again (if necessary) and always re-checked to ensure no more errors were found.

Multiple primaries
For the datasets that included a unique patient identification number, it was possible to check for multiple primary tumours following the IARC/IACR rules (IARC, 2004) especially defined for ICD-O-3 (see Chapter 5). The multiple primary check program can detect all the duplicate records that appeared during the reference period if the cancer registry submitted a complete dataset of the malignant cancer cases collected since the starting of registration; otherwise, some of the multiple tumours (generally those occurring at the beginning to the reference period) may not have been detected because of a lack of information on the earlier diagnoses.

Conversion into ICD-10
When no more errors remained, the incidence data were converted from ICD-O-3 to ICD-10. This ensured that the final ICD-10 codes used in the publication followed a standard ICD-O-3 to ICD-10 conversion program. When a dataset included an ICD-10 code, this was ignored in the tabulations. The ICD-O-3 to ICD-10 conversion program was written at IARC using the rules of the ICD-O-2 to ICD-10 conversion program, which was used for processing data for Volume VIII of CI5. The ICD-O-2 to ICD-10 conversion program was also developed at IARC, based on the *Conversion of neoplasms by topography and morphology from ICD-O-2 to ICD-9*. The terms deleted from ICD-O-3 (as indicated in Appendix 5) were first removed from the list of ICD-O-2. The conversion rules for the new ICD-O (M) codes (listed in Appendix 1) were defined by looking first at the most appropriate ICD-O-2 (M) code using the ICD-O-3 to ICD-O-2 conversion program, and then by applying the corresponding conversion rule to the ICD-O-3 (M) code. For example, the ICD-O-3 code M8174/3 *Hepatocellular carcinoma, clear cell type* is converted into ICD-10 following the rule that applies to the ICD-O-2 code M8170/3 *Hepatocellular carcinoma, NOS*. The ICD-O-3 codes M995_, M996_ (*myeloproliferative disorders*) and M998_ (*myelodysplastic syndromes*) that changed behaviour code from borderline (/1) to malignant (/3), and for which no ICD-10 code in the malignant 'C' category can be found, have been converted to the ICD-10 codes D45, D46_ and D47_ (i.e. non-malignant tumours). They are included and presented in the tables under the categories *MPD* and *MDS* respectively.

When a dataset was submitted with cases coded to ICD-9 or ICD-10 for topography and ICD-O-1 or ICD-O-2 for morphology, the series of conversion processes (Figure 6.2) and the final conversion from ICD-O-3 to ICD-10 produced some unexpected results and, for example, created new ICD-10 codes which were not originally recorded in the input file. For example, suppose the following combination of ICD-10 (T) and ICD-O-2 (M) was present:

ICD-10	ICD-O-2 (M)
C80	8640/3
Unknown primary	

The first conversion into ICD-O-2 (T+M) will produce the following output:

ICD-O-2 (T+M)	
C80.9	8640/3
Unknown primary	

The second conversion into ICD-O-3 (T+M) will produce the following output:

ICD-O-3 (T+M)	
C80.9	8640/3

Finally, the ICD-O-3 to ICD-10 conversion program used for the data processing will create the following ICD-10 code:

C62.9	Testis, NOS

Thus the final ICD-10 site becomes sex-specific and does not correspond to that provided in the original record. Generally, such change in ICD-10 codes will occur when the registry has not followed the rules in the ICD-O manuals; in the example above, a *Sertoli cell carcinoma* (M8640/3) should have been coded to testis (C62.9) if the site of the tumour was not specified (rule 8 of ICD-O-2 or rule H of ICD-O-3). It would also have occurred with other specific morphological diagnoses such as *malignant melanoma, regressing* (M8723/3) or *osteosarcoma* (M9180/3), which are automatically converted to an ICD-10 code for skin (C43._) or bone (C40._) cancer.

This example illustrates the importance of the primary conversion into ICD-O-3 and explains why the ICD-10 code provided by the cancer registry (if any) has not been used in the final tabulations. For certain morphological codes, the conversion is independent of topography. *Hepatocellular carcinoma* (M8170/3), for instance, is automatically converted to ICD-10 C22.0 irrespective of the ICD-O topography code (whether specific or unknown). The combination C34.9 (lung) plus M8170/3 will be converted to C22.0. This combination of topography and morphology is certainly incorrect. These types of potential errors are at the origin of the creation of the first version of the *IARC-CHECK* program. The validation of ICD-O-3 topography and histology combinations is an essential part of the data processing.

Miscellaneous conversions
In addition to the topography and the morphology codes, certain variables—sex, basis of diagnosis, ethnic group or race and dates—must be re-coded into a common schema following the instructions supplied by the cancer registry. When necessary, the basis of diagnosis variable was recoded following the IARC proposal (ICD-O-3, pg. 38). After the conversions into a common dictionary, the data corresponding to a registry are stored in the CI5 Vol. IX database, irrespective of whether they will be published in *Cancer Incidence in Five Continents Vol. IX*. All malignant neoplasms and non-malignant (except benign) neoplasms of the bladder are recorded in the database. These are stored as individual records containing the nine compulsory variables having topography and morphology coded to ICD-O-3 together with the corresponding ICD-10 code used for tabulation.

Mortality data
The mortality data are used for editorial purposes as an indicator of the completeness of registration, and are generally provided as a tabulation of ICD-9 or ICD-10 three-digit categories by sex and five-year age group, so that no validity check (except the basic combination of sex and site) can be performed. For some data sources, the original data were grouped by cancer sites or by wider age groups than the traditional five-year age groups, and had to be formatted before being handled by the series of editorial programs and stored in the CI5 Vol. IX database.

Population data
Cancer registries generally submitted population denominators for each individual year of the reference period, or for one year corresponding to the mid-year of the reference period. These are based on a census, which was carried out around the year

2000 in most of the countries. File editing and misinterpretation of the codes were the only source of errors discovered in the last three volumes. Whenever possible, the population data have been checked by comparing the age distribution with that from the previous volume. Unexpected change in the age structure or in the total population by year and sex were identified and sent to the registry for clarification. After this careful examination, the population files were formatted and stored in the CI5 Vol. IX database.

Conclusion

This protracted process for both cancer registry staff and the IARC secretariat took several months. However, it ensured a maximum level of data comparability and validity, but it is not in itself sufficient to ensure inclusion in the present volume. This depended upon other considerations of comparability and quality, as described in Chapter 5. The seven conversion programs, together with the latest version of the IARC-CHECK and the multiple primary check programs used in the incidence data process, have been published as a PC Windows™-based package, *IARCcrgTools*. This is available for free at the International Association of Cancer Registries (IACR) web site http://www.iacr.com.fr/. *IARCcrgTools* also includes the detailed definition of the edits and controls performed on each variable or within variables. All of the necessary programs to convert and check the data, then to create the tables were written in C++ and Java. The thousands of tables produced were generated in PostScript format and later converted into PDF files for publication.

The Editors would like to particularly thank Mr Mathieu Mazuir who performed most of the data entry and the data cleaning, Mr Eric Masuyer for his help in the data cleaning, and Mr Morten Ervik for his help in the data processing and his work in creating and managing the necessary tables.

Finally, the Editors would like to acknowledge the contributors who converted and checked their data prior to submission. This was very helpful and much appreciated.

References

Cooke, A.P., Parkin, D.M., Ferlay, J. CanReg4: Computer Software for Cancer Registries. Available at http://www.iacr.com.fr.

Ferlay, J., Burkhard, C., Whelan, S., Parkin, D.M. (2005). Check and Conversion Programs for Cancer Registries (IARC/IACR Tools for Cancer Registries). IARC Technical Report No. 42. Lyon, IARCPress.

Fritz, A., Ries, L. Conversion of Neoplasms by Topography and Morphology from the International Classification of Diseases for Oncology, Third Edition (ICD-O-3), to International Classification of Diseases for Oncology, Second Edition (ICD-O-2). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch.

Fritz, A., Percy, C.L., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D.M., Whelan, S. eds (2000). International Classification of Diseases for Oncology, third edition, Geneva, World Health Organization.

Parkin, D.M., Whelan, S.L., Ferlay, J., Teppo, L., and Thomas, D.B., eds (2002). Cancer Incidence in Five Continents, Vol. VIII. IARC Scientific Publications No. 155, Lyon, IARCPress.

Parkin, D.M., Chen, V.W., Ferlay, J., Galceran, J., Storm, H.H and Whelan, S.L. (1994). *Comparability and Quality Control in Cancer Registration*. (IARC Technical Report No. 19), Lyon, IARCPress.

Percy, C.L. ed. (1992). *Conversion of Neoplasms by Topography and Morphology from the ICD-0-2 to ICD-9 and the ICD-9-CM*. Washington, National Cancer Institute.

Percy, C.L, Van Holten, V. & Muir, C.S., eds (1990). International Classification of Diseases for Oncology, 2nd edition (ICD-O-2). Geneva, World Health Organization.

WHO (World Health Organization) (1957). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death* (Based on the Recommendations of Seventh Revision Conference, 1955), Geneva, World Health Organization.

WHO (World Health Organization) (1976). *International Classification of Diseases for Oncology* (ICD-O), Geneva, World Health Organization.

WHO (World Health Organization) (1977). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death* (Based on the Recommendations of Ninth Revision Conference, 1975), Geneva, World Health Organization.

WHO (World Health Organization) (1992). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death* (Based on the Recommendations of Seventh Revision Conference, 1990), Geneva, World Health Organization.