Chapter 6. Processing of data

J. Ferlay

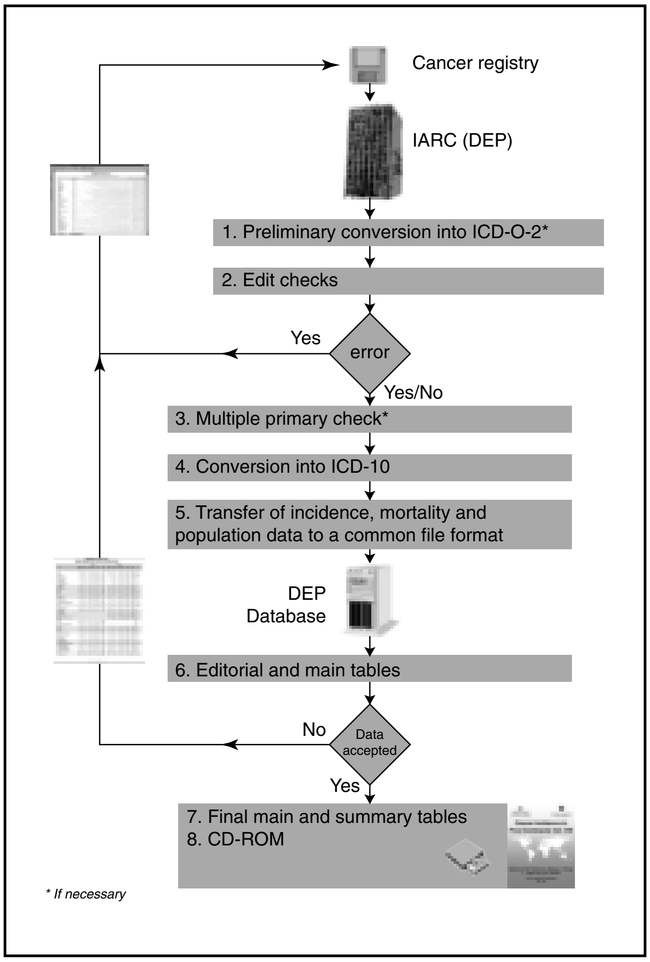
Cancer registry
IARC (DEP)
1. Preliminary conversion into ICD-O-2*
2. Edit checks
Yes
error
Yes/No
3. Multiple primary check*
4. Conversion into ICD-10
5. Transfer of incidence, mortality and population data to a common file format
DEP Database
6. Editorial and main tables
No
Data accepted
Yes
7. Final main and summary tables
8. CD-ROM
* If necessary

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

Around 235 cancer registries replied to the invitation to participate by submitting data for volume VIII. As usual, the data were accepted in any format and on any medium (paper forms, diskettes, electronic mail). This resulted in the manipulation of around 14 000 000 individual records, and the production of around 300 preliminary data-sets to be examined carefully by the editors (including different ethnic groups or time periods; see Table 6.1). A regular procedure for data validation and storage had been established and maintained in the Unit of Descriptive Epidemiology (DEP) of IARC. This data management process is designed to provide accurate data for a growing number of projects managed by the unit. *Cancer Incidence in Five Continents* is one of the major projects, and the preparation of the current volume is integrated with the regular work, which can be summarized by the diagram in Figure 6.1.

Data input processing

Incidence data

The incidence data were submitted as listings of individual anonymous cases with the following variables (minimum):

1. a registration number which 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
7. morphology of the tumour
8. behaviour of the tumour
9. basis of diagnosis

A description of all the codes used for these variables had to be provided with the data. However, it was not unusual that the code values did not match the description provided. In that case, the registry was asked for clarification and to provide the correct codes if necessary. This was particularly important when computing the percentage of histologically verified or DCO cases used for the

Table 6.1. Data for *Cancer Incidence in Five Continents Volume VIII*

Continent	Registries	Populations	Records (x1000)	Accepted in CI5	Populations
Africa	16	18	55	6	6
America, Central and South	18	18	300	11	11
America, North	30	66	5100	26	45
Asia	51	62	1500	43	50
Europe	108	108	6200	90	88
Oceania	12	17	900	10	14
Total	**235**	**289**	**14 055**	**186**	**214**

editorial process, because a misinterpretation of the basis of diagnosis code could give a false picture of the data quality. Many different coding schemes were used for tumour site and morphology as summarized in Table 6.2.

Table 6.2. Coding of information about the tumour

Topography	Morphology
ICD-10	No
ICD-7	MOTNAC (1968)
ICD-9	WHO/HS/CANC/24.1(1956)
ICD-9	MOTNAC (1968)
ICD-10	User-defined system
ICD-9 (3-digit)	MOTNAC (1968)
ICD-9 (3-digit)	ICD-O-1
ICD-9	ICD-O-1
ICD-9	ICD-O-2
ICD-10	ICD-O-2
ICD-O-1	ICD-O-1
ICD-O-1	ICD-O-Field Trial Edition (1988)
ICD-O-2	ICD-O-2

Conversion into ICD-O-2: The checking process using the *IARCtools* program requires the data to be coded by ICD-O-2. A great majority of the data-sets had to be converted into a full (topography and morphology) ICD-O-2 coding schema before they could be handled by the program. For registries using a mixture of ICD-9 or ICD-10 for topography, and ICD-O-1 or ICD-O-2 for morphology, specific programs were prepared and these proved to be particularly valuable in detecting incompatibilities between ICD-9 or ICD-10 codes and ICD-O-1 or 2 morphology and behaviour, which were transmitted back to the cancer registry for review and correction (Table 6.3).

Although the second edition of ICD-O gives clear instructions that behaviour codes /6 and /9 should not be used by cancer registries (page xxv of ICD-O-2), these codes appeared in many data-sets, 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) (Table 6.4).

Other difficulties that involved lengthy data processing should also be mentioned:

- Some registries did not correct their original files, so that all the corrections sent by fax or electronic mail had to be re-entered each time these registries re-submitted data.

- For registries submitting data coded to the main three-character categories of ICD-9 or ICD-10 only, a ‘dummy’ fourth digit had to be added to each ICD-9 or ICD-10 code. For the few registries that provided data coded to ICD-9 three-digit topography only, a special ‘main’ table had to be designed.

- For registries submitting records with ICD-9 or ICD-10 codes without histology, a ‘dummy’ ICD-O-1 or ICD-O-2 morphological code had to be assigned to each individual record to conform to the default data process. These registries are flagged using a ‘+’ indicating that no check regarding the validity of the diagnosis could be performed (see Chapter 5).

- Several data-sets included both ICD-O-1 and ICD-O-2, and sometimes both ICD-9 and ICD-10 codes depending on the year of incidence of the case. These data-sets had to be split into two (and for some of them into four) data-sets, each piece of the puzzle being converted using the appropriate program before a final full topography and morphology ICD-O-2 file could be handled by the check program.

- Some registries could not find an appropriate ICD-O-2 topography or morphology code and created their own codes. The corresponding cases had to be re-coded by hand to the most appropriate and valid ICD-O-2 code.

Checking: Once a data-set had been converted into ICD-O-2, or if it had been originally coded using ICD-O-2 codes, it was submitted to the *IARC-CHECK* program, which performed the following edits:

1. Code verification
- sex
- incidence and birth dates (if provided)
- ICD-O-2 topography and morphology

2. Consistency between items
- age versus birth/incidence dates
- sex versus site
- sex versus histology
- age versus site
- age versus histology

Table 6.3. Examples of unlikely ICD-10 site/ICD-O-2 morphology combinations

ICD-10		ICD-O second edition	
C82._	Non-Hodgkin lymphoma	Any ICD-O (M) code less than 9590	
C81._	Hodgkin disease	Any ICD-O (M) code less than 9650	
C46._	Kaposi sarcoma	ICD-O (M) not 9140/3	
C91.0	Acute lymphoid leukaemia	9823/3	Chronic lymphocytic leukaemia
C81.9	Hodgkin disease, NOS	9590/3	Non-Hodgkin lymphoma, NOS
C43.9	Melanoma of skin, NOS	8090/3	Basal cell carcinoma
C34.9	Lung (primary cancer)	8140/6	Adenocarcinoma, metastatic
C53.9	Uterine cervix, malignant	8070/2	In situ tumour

Table 6.4. Examples of unlikely combinations of items

Sex	Age	Topography	Morphology	Basis of diagnosis
Male		C76.7	8930/3	
	30	C61.9	8140/3	
	60	C64.9	8960/3	
		C61.9	9140/3	
		C34.9	8170/3	
		C42.1	9827/3	Non-microscopic verification

- site versus histology
- basis of diagnosis versus histology

Registries submitting data for Volume VIII had been invited to run their own data through the *IARC-CHECK* program before submission, and a number of contributors did so. For the other registries, all errors or unlikely or rare combinations of items were sent back to the cancer registry for verification. The amendments or new files resubmitted were then incorporated, converted (if necessary) and always checked again to ensure that no more errors were found. This long and tedious process for both cancer registry and DEP staff took several weeks or months; however, it ensured a maximum level of data comparability and validity. This validation process was 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.

Multiple primaries: When a data-set incorporated an identification number which was a *patient* identification number, it was possible to check for multiple primary tumours following the IARC/IACR rules (IARC, Lyon, 1994) (Figure 6.2). In this case, the data file was first sorted on the identification number, and within the identification number, by ascending incidence date. All the records concerning the patient were then passed through the following algorithm to detect multiple tumours or true duplicate registrations:
Suppose there were many records for the same patient:

- T(i) being the topography (three digits of ICD-10)
- M(i) being the morphology
- TG(k) and MG(k) being the groups of topography and morphology considered to be different (Parkin *et al.*, 1994, pp. 3 and 4)

This program can detect all the duplicates which appeared during the period *only* if the cancer registry has submitted its complete data-set, including the years before the period for the current volume. Otherwise, some of the multiple tumours (generally those which occurred at the beginning to the current period) might not be detected because of lack of information on the prevalent cases.

Conversion into ICD-10: When no more errors remained, the incidence data were converted from ICD-O-2 to ICD-10. This ensured that *the final ICD-10 codes used in the publication followed a standard ICD-O-2 to ICD-10 conversion program.* When a data-set was submitted coded to ICD-10, the series of conversion processes produced some unexpected results and, for example, created 'artificially' new ICD-10 codes which were not originally recorded in the input file. Suppose the following combination of ICD-10 (T) and ICD-O-2 (M) was present:

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

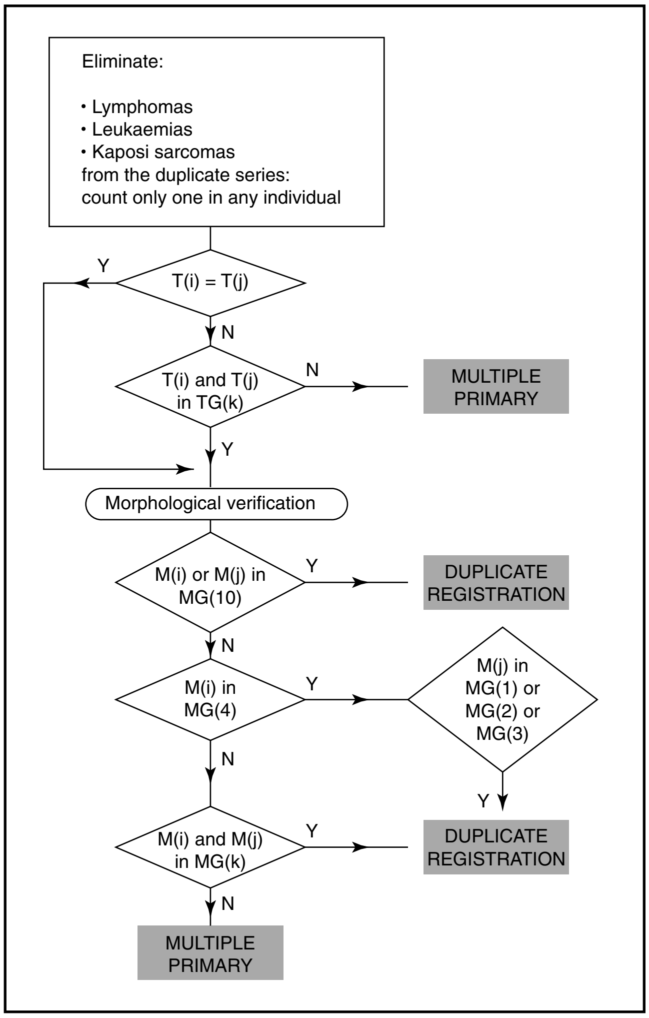

Figure 6.2. Processing for detection of multiple primaries

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

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

Finally, the ICD-O-2 to ICD-10 conversion program used for the data processing will produce the following ICD-10 code:

C62.9	Testis, NOS

so that the final ICD-10 site becomes sex-specific and does not correspond to that provided in the original record. Generally, such code changes 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). But it would also have occurred with other specific morphological diagnoses such as basal cell carcinoma (M8090) or osteosarcoma (M9180), which would be converted to an ICD-10 topography code for skin or bone cancer. This explains why some cancer registries that submitted their data coded to ICD-10 found differences between their tabulations and those produced by the *Cancer Incidence in Five Continents* process. In addition to the potential errors described above, this is another reason why IARC strongly recommends the use of ICD-O for coding morphology *and* topography.

All the conversion and check programs used in the data-entry process have been published as a PC Windows™ based package IARCtools (Ferlay, 1997), available free on the Internet at http://www-dep.iarc.fr/resour/software/iarctools.htm. A new version that will work with ICD-O-3 codes will be available in 2003.

Miscellaneous conversions: Before being loaded into the DEP database, each variable within a data file (sex, basis of diagnosis, ethnic group or race, dates, etc.) had to be re-coded into a common schema, following the instructions given by the cancer registry. For example, the basis of diagnosis code was recoded following the IARC schema (ICD-O-2, page xxxix) if necessary. The DEP database contains all the incidence data-sets received, irrespective of whether or not they are published in *Cancer Incidence in Five Continents*. The incidence data are stored as individual records and, as described above, checked and coded to ICD-O-2. The database contains currently more than thirty million records that can be easily converted to any other classification system for different collaborative studies.

Mortality data

The mortality data used for editorial purposes 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. Depending on the source, the original data might be grouped by cancer site or by wider age-groups than the traditional five-year age bands, so that they had to be formatted before being handled by the series of editorial programs, and stored in the DEP database.

Population data

Cancer registries generally submitted population denominators corresponding to the mid-year of the period of interest, based on a census or survey. However, some registries provided data from two or more censuses, or estimates for years outside the period for which data on cancer cases were submitted, so that a more precise estimate of the person-years at risk could be calculated. The population figures were then appropriately formatted and loaded into the DEP database for future use.

Output data processing

Using the DEP database, the production of the volume was quite fast. The data corresponding to the period of the registries accepted for publication were retrieved and converted into Cancer Incidence Five Continents morphological groups (see Chapter 4). The resulting files are recorded on the CD-Rom in a tabulated format (see Chapter 7) and also used by the *CI5VIII* software (see Chapter 7). The editorial and the final tables presented in the book were produced using specially designed programs running in batch mode.

The complete data processing was performed using a standard PC running Windows™2000 with a sufficient amount of disk storage (40 gigabytes). All the necessary programs to convert and check the data, then to create the tables were written in C++. The thousands of tables produced were generated in PostScript format for printing purposes, and after validation, converted into PDF files prior to publication. The DEP database is located on a separate server that runs Windows™ SQL Server2000.

Conclusion

The author would like to thank all the persons involved in the validation procedure for their patience and their unfailing help, and to particularly acknowledge the contributors who checked their data using *IARCtools* program before submission. This was very helpful and much appreciated.

References

Ferlay, J. (2002) *Cancer Incidence in Five Continents* VIII (CancerBase 6), Lyon, IARC

Manual of Tumor Nomenclature and Coding (1951), New York, American Cancer Society

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

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, DC, National Cancer Institute

Percy, C.L., Berg, J.W. & Thomas, L.B., eds (1968) *Manual of Tumor Nomenclature and Coding*, Washington, DC, American Cancer Society

Percy, C.L. & Van Holten, V., eds (1988) ICD-0, *International Classification of Diseases for Oncology, Field Trial Edition* (developed by working party coordinated by IARC, Lyon)

WHO (World Health Organization) (1956) *Histological classification of neoplasms*, (WHO/HS/CANC/24.1), Geneva

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

WHO (World Health Organization) (1990) *International Classification of Diseases for Oncology*, Second Edition (ICD-O), Geneva

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

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

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