Chapter 6: Data processing

Jacques Ferlay, Eva Steliarova-Foucher, Sébastien Antoni, and Eric Masuyer

The call for data for Volume X of *Cancer Incidence in Five Continents* (CI5) was launched in September 2011, as described in Chapter 1. The call included detailed instructions about the content and format of the material to be submitted, and was disseminated to 596 identified cancer registries worldwide and posted on the website of the International Association of Cancer Registries (IACR). Registries wishing to submit data for inclusion in CI5 Volume X were asked to provide cancer incidence and mortality data, population data, a 350-word introductory text (narrative), a completed questionnaire, a coding schema, and other relevant information.

DATA FLOW

For this volume, CI5 data were collected semi-automatically for the first time. Registries were asked to submit all material via the Registries Portal, at a secure website (https://cinportal.iarc.fr/). The portal was developed as part of the EUROCOURSE project (http://www.eurocourse.org/), in collaboration with the National Cancer Registry Ireland and the Section of Cancer Surveillance (CSU) at IARC (Steliarova-Foucher et al., 2014). The portal is equipped with a series of programs that enable the automatic exchange of information between the cancer registries and CSU. Based on each registry's access credentials and the submitted data file type, uploaded files were automatically named and stored in an organized system of folders on an internal server at IARC. The submitting registry and the designated CSU staff members were notified of each submission by an automatically generated email. Throughout the process, the registries could review their submissions and manage their uploaded files at any time.

The portal was also used for communication between IARC and potential contributors during the editorial process. Requests for data correction or supplementary information, as well as decisions about registries' inclusion in the volume, were communicated through the portal. The relevant files were uploaded to the registry-specific *Feedback* section of the portal, where the registries could retrieve the files after being notified by an automatically generated email. The registries could then submit their responses and any revised or supplementary data in the same way as their initial material. A log of the files' movements on the IARC server was monitored by CSU.

Most of the invited registries submitted and exchanged data through the secured Registries Portal, but a few registries submitted zipped files by email. A schematic representation of the overall flow of data and processing steps is shown in Fig. 6.1.

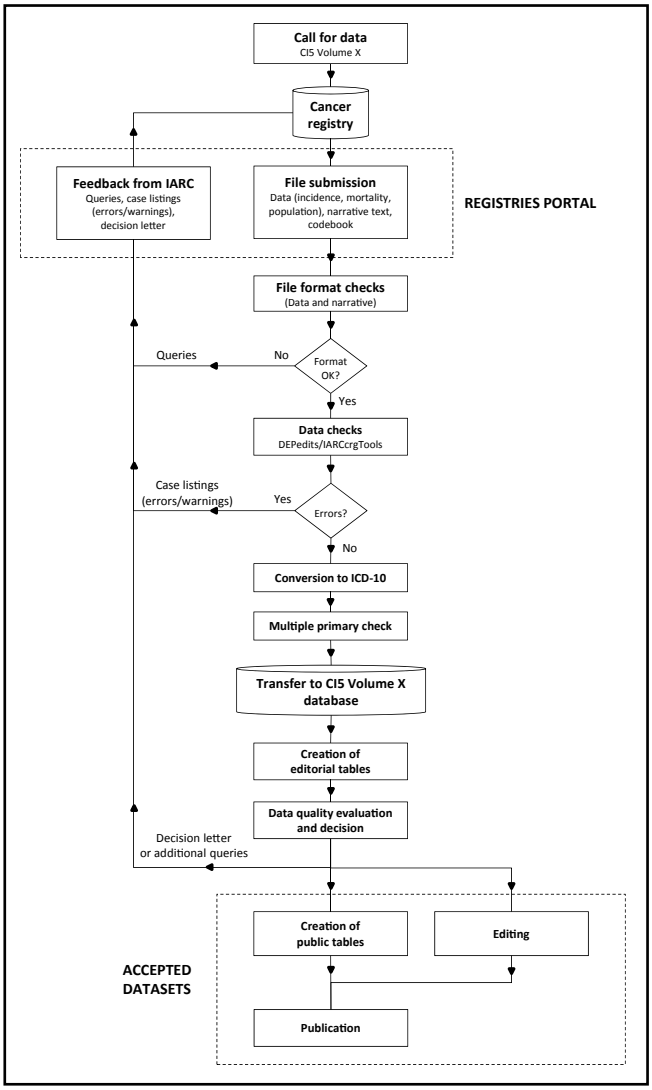

Fig. 6.1. A schematic representation of the overall flow of data and processing steps in the creation of Volume X of *Cancer Incidence in Five Continents*.

DATA PROTECTION

All raw, individual-level data collected for CI5 Volume X were stored on a secure protected server at IARC, to which only a limited number of selected CSU staff members had access. These data will not be used for any other purpose or transferred to any third party without the registries' explicit permission.

DATA PROCESSING

A total of 370 cancer registries submitted data in response to the invitation to participate in CI5 Volume X (see Chapter 1). Although the preferred file formats were specified in the study protocol, data were received in several electronic formats (text files, spreadsheets, database files, etc.), with varying layouts. The first step of data processing therefore included a quick check of the files' contents (sometimes resulting in a request for additional material), as well as some reorganization and formatting.

About 40 million individual cancer records were received and processed by IARC. Coupled with mortality files, preliminary datasets representing 545 populations (including various ethnic groups) were produced and reviewed by the editors (see Chapter 5). All submitted data were processed and checked by IARC using automated in-house processes based on standard measures of data quality.

The IARC software packages DEPedits and IARCcrgTools (Ferlay et al., 2005; available from http://www.iacr.com.fr/) were used to check and convert the data. All programs used to process, check, and convert the data and create the tables were written in Stata and C++. The thousands of tables produced during the editorial process for publication online and in this volume were generated in PostScript format and converted to PDF.

Incidence data

Registries were asked to submit their incidence data as individual anonymized case listings including all malignant tumours and non-malignant tumours of the bladder, collected for the longest period possible, and to include incident diagnoses for a minimum of 3 consecutive years within the period of 2003–2007. Each record contained at least the following variables:

1. A registration number that uniquely identified the patient
2. Sex
3. Ethnic group or race (optional)
4. Birth date and/or age at incidence date
5. Incidence date
6. Tumour site (topography)
7. Tumour morphology
8. Tumour behaviour
9. Most valid basis of diagnosis.

Descriptions of the codes used for each variable were also required. However, it was not unusual for code values not to match the descriptions provided, or for coding information to be missing. In such cases, the registry was asked for clarification and to provide the correct codes if necessary. This was particularly important for calculating the percentage of microscopically verified or death-certificate-only (DCO) cases, for evaluating the important indicators of data quality influencing the decision to include a dataset in the volume, and for determining the potential designation of data with an asterisk.

CONVERSION TO ICD-O-3

Although 96% of the registries (356/370) submitted data already coded to ICD-O-3 (Fritz et al., 2000), 16 datasets had to be converted to ICD-O-3 before they could be processed by the program. The alternative coding systems used were ICD-O-1 (WHO, 1976), ICD-O-2 (Percy et al., 1990), and sometimes combinations of ICD-9 (WHO, 1977) or ICD-10 (WHO, 1992) topography with ICD-O morphology. Conversion from ICD-9 or ICD-10 combined with ICD-O required partitioning of the original file into two or more files; each was then converted separately using the appropriate program, then merged. These preliminary conversions helped to detect incompatibilities between ICD-9 or ICD-10 codes and the ICD-O system. Any incompatible records were sent back to the registry for review and correction.

Although ICD-O-3 clearly states that behaviour codes /6 and /9 should not be used by cancer registries (ICD-O-3, p. 27), these codes did appear in a few datasets, casting doubt on the accuracy of the corresponding topography codes. Usually, the provided topography code was assumed to represent the site of the primary tumour. When this was evidently not the case (e.g. carcinomas in lymph nodes or bone), a list of such cases was sent back to the registry with a request for clarification. As a last resort, these cases were recoded to topography C80.9 (primary site unknown), the provided morphology code was retained, and the behaviour code was set to /3.

CHECKING

All datasets with complete ICD-O-3 coding of tumour site, morphology, behaviour (and optionally grade), and basis of diagnosis were run through the IARC-CHECK program included in the IARCcrgTools package, which performed the following checks:

1. Code verification
   - Sex
   - Incidence date (and birth date, if provided and complete)
   - ICD-O-3 topography, morphology, and behaviour

2. Consistency between items
   - Age versus birth and 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 Volume X were invited to run their data through the IARC-CHECK program before submission, and a large number of contributors did so. The datasets of registries using CanReg software (available from http://www.iacr.com.fr/) were checked using the equivalent built-in functionalities. All datasets were rechecked by IARC staff. Any errors or unlikely or rare combinations of items were sent back to the registry for verification, unless they were already flagged as double-checked. The received corrections and resubmissions were then consolidated, converted if necessary, and rechecked to ensure that no further errors were found. More than

one cycle of data validation was required for many of the datasets.

MULTIPLE PRIMARIES

All records included a unique patient and tumour identification number, so it was possible to check for multiple primary tumours occurring in the same patient using the multiple primary check program included in the IARCcrgTools package. The software lists all sets of tumours recorded for a single patient that should be considered a single primary tumour according to the IARC/IACR rules specifically defined for ICD-O-3 (IARC, 2004). The longer the time period for which data were submitted, the more complete the identification of multiple primary tumours in the reference period of 2003–2007.

CONVERSION TO ICD-10

When no errors remained, the incidence data were converted from ICD-O-3 to ICD-10, to ensure that the ICD-10 categories resulted from the same conversion process (ICD-O-3 to ICD-10) for all cancer registries. The ICD-O-3 to ICD-10 conversion program was written at IARC and is based on the rules defined in *Conversion of Neoplasms by Topography and Morphology from ICD-O-2 to ICD-10* by Percy (1998). In summary, each new ICD-O-3 morphology code (as listed in Appendix 1 of ICD-O-3) was converted first to the closest ICD-O-2 morphology code using the ICD-O-3 to ICD-O-2 conversion program (Fritz and Ries, 2001), and then the corresponding ICD-O-2 to ICD-10 conversion rule was applied. For example, the ICD-O-3 code M8174/3 (Hepatocellular carcinoma, clear cell type) was converted to the ICD-10 code C22.0, following the rule that applies to the ICD-O-2 code M8170/3 (Hepatocellular carcinoma, not otherwise specified [NOS]).

The conversion rules strictly follow the ICD-10 coding rules expressed in the instruction manual of ICD-10 Volume 2 and the alphabetical index of ICD-10 Volume 3. For example, the combination of unknown primary site (ICD-O-3 topography code C80.9) and fibrosarcoma, NOS (ICD-O-3 morphology code 8810/3) was converted to ICD-10 code C49.9 (Connective and other soft tissues, NOS; see ICD-10 Volume 2, p. 74). The ICD-O-3 codes M995_, M996_ (myeloproliferative disorders [MPD]), and M998_ (myelodysplastic syndromes [MDS]), for which no ICD-10 code in the malignant C category exists, were converted to the ICD-10 codes D45, D46_, and D47_ (i.e. non-malignant tumours), respectively, and are included and presented in the tables under the categories MPD and MDS.

When a dataset was submitted with cases coded to ICD-10 for topography and ICD-O-2 for morphology, the conversion process sometimes produced ICD-10 codes different than those originally provided in the submitted file, as shown in Table 6.1.

In the example shown in Table 6.1, the final ICD-10 code is sex-specific and differs from the code provided in the original record. Generally, such changes in ICD-10 codes occurred when a registry did not strictly follow the rules in the ICD-O manuals; in the example shown in Table 6.1, 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). The ICD-10 code for unspecified site was recoded to skin (C43.9) or bone (C41.9) if the morphological diagnosis was malignant melanoma, regressing (M8723/3) or osteosarcoma (M9180/3), respectively. This example explains why the original ICD-10 codes provided (if any) were not used in the final tabulations.

For certain morphological codes, the conversion was independent of topography. For example, hepatocellular carcinoma (M8170/3) was automatically converted to the ICD-10 code C22.0, irrespective of the ICD-O topography code (whether specific or unknown). Thus, the combination of ICD-O-3 topography code C34.9 (lung, NOS) and morphology code 8170/3 was converted to ICD-10 code C22.0, because the original combination of topography and morphology was obviously incorrect. It was the detection of this kind of error that inspired the creation

Table 6.1. Conversions of datasets with cases coded to ICD-10 for topography (T) and ICD-O-2 for morphology (M)

Classification	Original coding	First conversion	Second conversion	Third conversion
ICD-10	C80 – Unknown primary			C62.9 – Testis, not otherwise specified (NOS)
ICD-O-2 (T)		C80.9 – Unknown primary		
ICD-O-2 (M)	8640/3 – Sertoli cell carcinoma	8640/3 – Sertoli cell carcinoma		
ICD-O-3 (T)			C80.9 – Unknown primary	
ICD-O-3 (M)			8640/3 – Sertoli cell carcinoma	

of the first version of the IARC-CHECK program, and these errors were detected during the checking process described earlier in this chapter.

MISCELLANEOUS CONVERSIONS

In addition to tumour topography and morphology, certain other variables (sex, basis of diagnosis, ethnic group or race, and dates) were also recoded to a common system according to the descriptions supplied by the registries. When necessary, the basis of diagnosis variable was recoded following the ICD-O-3 scheme, and for a few registries, the incidence dates required conversion to the Gregorian calendar. After conversion to a common dictionary, the incidence data corresponding to each registry were added to the CI5 Volume X database.

Mortality data

Together with their data on cancer cases, the registries were asked to provide official cancer mortality data for the reference period (2003–2007), ideally for each calendar year of the period. For the national cancer registries, mortality data were extracted from the WHO Mortality Database (http://www.who.int/healthinfo/statistics/mortality_rawdata/en/) by IARC staff. The mortality data were used for editorial purposes as an indicator of the completeness of registration. Because the data were generally provided in tabular form for the available ICD-9 or ICD-10 three-digit categories by sex and 5-year age group, only checks for the validity of the ICD code and the combination of sex and site were performed. The data provided by some registries were grouped into wider cancer sites or age groups than conventionally used, and therefore had to be reformatted before being processed by the series of editorial programs and added to the CI5 Volume X database. Some registries supplied mortality data based on the cancer registry dataset. Such data were not considered by the volume editors to represent official mortality data.

Population data

The registries were required to submit population denominators for each individual year of the reference period, by sex and 5-year age groups. In the absence of corresponding data sources, a population denominator for a single central year of the reference period was accepted. The population data were checked first by careful examination of the data file, then by comparing the age distribution with that from the previous CI5 volume, if available. Unexpected changes in the age structure or in the total population by year and sex were identified and queried. After examination, the population files were formatted and added to the CI5 Volume X database.

THE CI5 VOLUME X DATABASE

The CI5 Volume X database contains all the incidence, mortality, and population datasets supplied by the registries and checked by IARC for the project, irrespective of whether they were ultimately selected for inclusion in the volume itself. Incidence data were supplied for all malignant neoplasms, as well as non-malignant (except benign) neoplasms of the bladder. The data are stored as a listing of individual records (without patient identification numbers, which are no longer necessary) containing the eight compulsory variables, with topography and morphology coded to ICD-O-3 together with the corresponding ICD-10 code used for tabulation. The database is hosted and maintained on a protected server at IARC, with access restricted to identified CSU staff members.

CONCLUSION

The complete process of data checking and validation conducted by IARC in collaboration with the cancer registries took several months. The help provided by those contributors who converted and checked their data before submission was greatly appreciated. Although prompt data provision is of the utmost importance, this importance is counterbalanced by the necessity of validating and ensuring the comparability of the global cancer incidence data. Online publication of the data ahead of the printed version provided earlier public access to the results.

The data processing methods described in this chapter resulted in the standardization of the information provided, which allowed the CI5 editors to compare datasets within large defined geographical regions, as described in Chapter 5. The CI5 data validation processes contributed substantially to the overall quality and comparability of the data from all submitting registries, as well as to data harmonization, with the benefit extending beyond this publication.

REFERENCES

Ferlay J, Burkhard C, Whelan S, Parkin DM (2005). Check and Conversion Programs for Cancer Registries (IARC/IACR Tools for Cancer Registries). IARC Technical Report No. 42. Lyon: International Agency for Research on Cancer.

Fritz A, Percy CL, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al., editors (2000). International Classification of Diseases for Oncology. 3rd ed. (ICD-O-3). Geneva: World Health Organization.

Fritz A, Ries L (2001). Conversion of Neoplasms by Topography and Morphology from the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) to the International Classification of Diseases for Oncology, Second Edition (ICD-O-2). Bethesda, MD: National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program, Cancer Statistics Branch.

IARC (2004). International Rules for Multiple Primary Cancers ICD-O Third Edition. Internal Report No. 2004/02. Lyon: International Agency for Research on Cancer.

Percy CL, editor (1998). Conversion of Neoplasms by Topography and Morphology from the International Classification of Diseases for Oncology, Second Edition (ICD-O-2) to the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10). Bethesda, MD: National Cancer Institute, National Institutes of Health.

Percy CL, Van Holten V, Muir CS, editors (1990). International Classification of Diseases for Oncology. 2nd ed. (ICD-O-2). Geneva: World Health Organization.

Steliarova-Foucher E, O'Callaghan M, Ferlay J, Masuyer E, Rosso S, Forman D, et al. (2014). The European Cancer Observatory: a new data resource. Eur J Cancer. http://dx.doi.org/10.1016/j.ejca.2014.01.027 PMID:24569102.

WHO (1976). International Classification of Diseases for Oncology. 1st ed. (ICD-O-1) Geneva: World Health Organization.

WHO (1977). International Classification of Diseases. 1975 revision (ICD-9). Geneva: World Health Organization.

WHO (1992). International Statistical Classification of Diseases and Related Health Problems. 10th revision (ICD-10). Geneva: World Health Organization.