Chapter 6: Data processing

Murielle Colombet, Aude Bardot, and Jacques Ferlay

The call for data for Volume XII of *Cancer Incidence in Five Continents* (CI5) was launched in July 2021. The call included detailed instructions about the content and format of the material to be submitted and was disseminated to all identified cancer registries worldwide and posted on the IACR website. Registries that wished to submit data for inclusion in CI5 Volume XII 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

Registries were asked to submit all material via the IARC Registries Portal, at a secure website (https://cinportal.iarc.fr/). The portal enables the automatic exchange of information between the cancer registries and the CI5 Secretariat at the Cancer Surveillance Branch (CSU) at IARC. Based on each registry's access credentials and the submitted data file types, 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, table revision, 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 staff members.

A schematic representation of the overall flow of data and processing steps is shown in Fig. 6.1.

DATA PROTECTION

All raw, individual-level data collected for CI5 Volume XII 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.

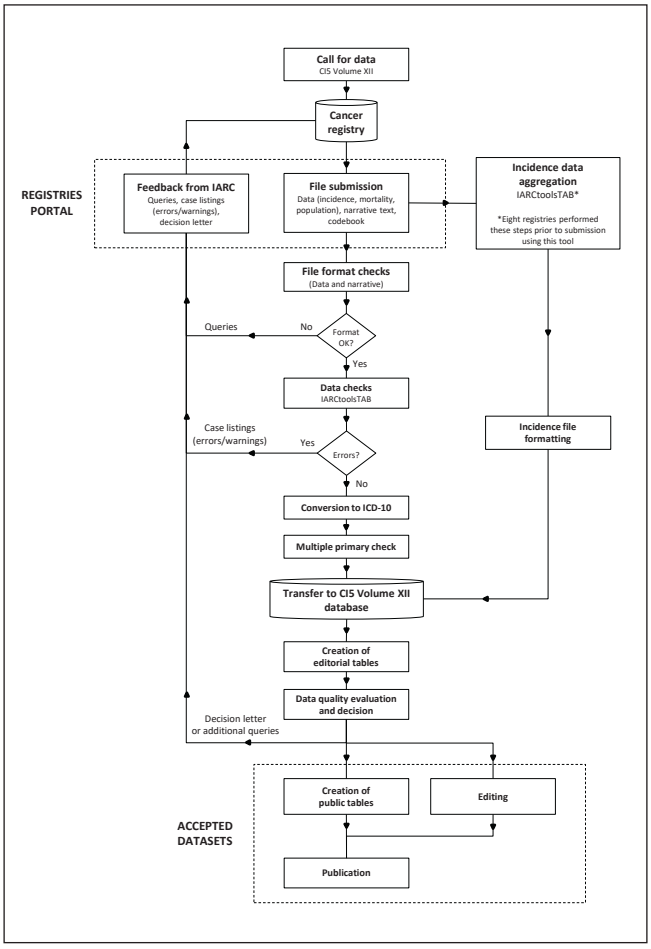

Fig. 6.1. A schematic representation of the overall flow of data and processing steps in the creation of Volume XII of *Cancer Incidence in Five Continents*.

DATA PROCESSING

A total of 675 cancer registries submitted data in response to the invitation to participate in CI5 Volume XII. 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. Therefore, the first step of data processing included a quick check of the files' contents (sometimes resulting in a request for additional material), as well as some reorganization and formatting.

About 149 million individual cancer records were received and processed by IARC. The largest dataset, with about 32 million records, was supplied by the

National Program of Cancer Registries (NPCR) in the USA. Coupled with mortality files, preliminary datasets representing 813 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 package IARCtoolsTAB (Colombet et al., 2020) was used to check and convert the data. This tool was also made available to population-based cancer registries that wished to supply their incidence data for CI5 Volume XII in an aggregated format. Eight registries provided aggregated data using this tool (cancer registries of Denmark, Finland, Iceland, Japan, The Netherlands, Norway, Sweden, and UK, England). All programs used to process, check, and convert the data and create the tables were written in Stata, R, and C++.

Incidence data

Registries were asked to submit their incidence data as individual anonymized case listings including all malignant tumours and non-malignant tumours (except benign tumours) of the bladder and of the central nervous system, collected for the longest period possible, and to include incident diagnoses for a minimum of 3 consecutive years within the period of 2013–2017. 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 almost all registries submitted data already coded to ICD-O-3 (Fritz et al., 2000) or to the 2011 revision of ICD-O-3 (WHO, 2013), three datasets had to be converted to ICD-O-3 before processing. The alternative coding system used in the three population-based cancer registries was a combination of ICD-10 (WHO, 1992) topography with ICD-O morphology. Conversions from ICD-10 combined with ICD-O required a two-step process: conversion from ICD-10 to ICD-O-2 and then to ICD-O-3. These preliminary conversions helped to detect incompatibilities between ICD-10 codes and the ICD-O system. Any incompatible records were sent back to the registry for review and correction.

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 (Ferlay et al., 2005) included in the IARCtoolsTAB package, which performed the following checks:

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

2. Consistency between items
   - Age versus birth and incidence dates
   - Chronology between birth and incidence dates
   - Sex versus site
   - Sex versus histology
   - Age versus site
   - Age versus histology
   - Site versus histology
   - Basis of diagnosis versus histology
   - Behaviour versus site
   - Behaviour versus histology.

Registries submitting data for CI5 Volume XII were invited to process their data through the IARC-CHECK program before submission, and most did so. The datasets of registries that used CanReg software (available from http://www.iacr.com.fr/) were checked using the equivalent built-in functionalities. All datasets were rechecked by CSU staff members. 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 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 IARCtoolsTAB 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) and modified to take into consideration the new terms that appeared in the 2011 revision (see Chapter 3). The longer the time period for which data were submitted, the more complete the identification of multiple primary tumours within the reference period (2013–2017).

CONVERSION TO ICD-10

When no errors remained, the incidence data were converted from the first (2011) revision of ICD-O-3 to ICD-10 (2010 version), 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 2011 revision of ICD-O-3 affected the morphology numerical list only, introducing new terms that have appeared in the recent literature, particularly in the "Lymphoma and leukaemia" group. Therefore, a new conversion table from the first (2011) revision of ICD-O-3 to ICD-10 (2010 version) for the ICD-O-3 morphology codes 9590/3 to 9992/3 was developed (see Chapter 3).

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 M9950/3, M9960–9965/3, M9971/3, and M9975/3 (myeloproliferative disorders [MPD]), and M9980–9983/3, M9985–9989/3, M9991/3, and M9992/3 (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 (see Chapter 3).

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.

Mortality data

Together with data on cancer incidence, registries were asked to provide official cancer mortality data for the reference period (2013–2017), ideally for each calendar year of the period. For the national cancer registries, mortality data were extracted from the WHO Mortality Database (https://platform.who.int/mortality/) by CSU staff members. 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-10 (and sometimes ICD-9) three-digit categories by sex and 5-year age groups, 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 XII database. Some population-based cancer registries supplied mortality data based on the cancer registry dataset. Such data were not considered by the volume editors to represent official mortality data (see Chapter 5).

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 XII database.

THE CI5 VOLUME XII DATABASE

The CI5 Volume XII 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. The data (i.e. the total number of cases, the microscopically verified cases, and the DCO cases) are stored in aggregated format by 18 age groups, sex, calendar year, and 347 disease entities. The database is hosted and maintained on a protected server at IARC, with access restricted to identified CSU staff members.

CONCLUSIONS

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 published volume provided earlier public access to the results.

The data processing methods described in this chapter resulted in the standardization of the information provided, which enabled 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

Colombet M, Bray F, Ferlay J (2020). IARCtoolsTAB package: a federated solution to registry data submission. Lyon: International Agency for Research on Cancer.

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.

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

WHO (2013). International Classification of Diseases for Oncology. 3rd ed. (ICD-O-3), 1st revision. Geneva: World Health Organization.