CHAPTER 4.

Sources of information for the population-based cancer registry

A key feature of the PBCR is the use of multiple sources of information on cancer cases in the target population. This facilitates the identification of as many as possible of the cases diagnosed among the residents of the registry area. It does not matter if information on the same cases is received from several sources (indeed, as described in Chapter 5, this feature of a PBCR may be used to evaluate its success in case finding). Registry procedures allow identification of the same cancer case from different sources (while avoiding duplicate registrations); this is a built-in feature of the CanReg5 software (see Annex 1).

1. Sources of information on cancer cases

The sources can be grouped into three broad categories, each of which is discussed below:

- hospitals
- laboratories
- death certificates.

1.1 Hospital sources

The registry should attempt to identify all cancer cases that are diagnosed or treated in hospitals or clinics in the registry area. The institutions concerned will vary depending on location, but it is important to identify and enumerate them all, and the likely number (and type) of cancer patients seen in each. If there are special cancer treatment facilities (medical/surgical oncology, radiotherapy), their contribution to the registry is essential. Often, such services maintain a register of cases diagnosed, treated, or under follow-up.

Most other hospital services will see cancer patients, although the proportion of cases that are malignant disease will vary depending on the specialty. If there is a hospital information system, from which patients plus their diagnoses can be abstracted, the registry will use this as the primary case-finding mechanism. Even without a computerized hospital information system, the medical records department may maintain manual indexes of hospital discharges, which can be sorted by diagnosis. When there is no central information system, the work of the registry is more laborious and may involve visits to individual clinical services.

Private hospitals or clinics tend to be smaller than the larger public hospitals, and may not have specialist treatment facilities for cancer. Nevertheless, they may be important to include among the data sources, if identification of cancer patients

among their clientele is relatively easy. Confidentiality issues (real or imagined) with respect to collaboration with the cancer registry may be raised by the owners.

Hospice and palliative care services are very important sources. The great majority of their clients are cancer patients, documentation of diagnosis is usually good, and follow-up until death is the norm (indeed, the purpose).

1.2 Laboratory services

The pathology laboratory is a key – indeed, essential – source of data. For most cancer patients, the definitive diagnosis is based on histology (although the proportion of cases for which the tumour is examined by the pathologist depends on the site/type of cancer). Pathology laboratories always keep a record of their work in the form of a register – often as a computerized database, but even paper registers are easy to scan for cancer diagnoses. However, the laboratory will often be dependent on the request form, which accompanies the specimen, for information on the cancer patient. These, in turn, may contain inadequate information or be badly completed – especially with respect to place of residence. This variable is essential to the PBCR, and special effort is needed to find the information for cases found via the laboratory.

Other laboratory services are less fruitful sources, although clinical haematologists (rather than pathologists) are generally responsible for examination of bone-marrow specimens (and hence for diagnosis of haematological malignancies). Among the medical imaging services, only magnetic resonance imaging (MRI) and computed tomography (CT) scans have a high enough yield of cancer cases to be worth considering as sources of data. Their utility depends on the ease with which cancer cases can be identified among the lists of patients examined.

1.3 Death certificates

Information on persons dying from (or with) cancer is a very important source of case data for the registry. This information may be from civil registration systems (where "cause of death" is recorded by a medical practitioner on a death certificate), even if this process is incomplete (in the sense that not all deaths are certified). The correct assignment and coding of cause of death are often a problem in civil registration systems in LMICs. In many lower-income countries, death registration is confined to deaths in hospital (with no medical certification for deaths occurring at home); even these limited data should be exploited by the registry.

Identifying individuals dying from (or with) cancer serves three purposes for the registry:

- It allows identification of cancer cases that had been "missed" by the data collection system.
- It allows the death of registered cancer cases to be recorded (used in calculation of survival).
- Knowledge of the numbers of cases first notified via a death register provides one method for estimating the completeness of cancer registration.

2. Data collection

Traditionally, a distinction is made between "passive" collection of data (relying on health workers to complete notification forms and forward them to the registry) and "active" methods, whereby staff of the cancer registry visit the various sources to identify and abstract the relevant information. Registration that relies entirely on the diligence and goodwill of others to do the work of abstraction of information on cancer cases is never successful. Nevertheless, most registries use a mixture of methods, and although active case finding remains the norm, the development of computerized health information systems provides some scope to use electronic databases for case finding.

With an increasing number of computerized data sources available, cancer registries are sometimes put under pressure to abandon their traditional modes of operation. Whereas in the long term registries should develop a strategy to move from paper to digital data sources, it is a misconception to believe that cancer registry data can be automatically derived from the health information system. Regardless of the data sources and data collection methods used, skilled cancer registry staff are required to produce high-quality incidence data. In some LMICs, the person-time available for cancer registration allows only for routine data processing and production of incidence data. Making use of data from health information systems could enable such registries to spend less person-time on data entry and allocate more time to quality control, data analysis, and possibly research.

3. Variables collected by the registry

Cancer registries set out to record data for a set of variables on each cancer case. There is a uniform tendency, when a cancer registry is planned, to aim for too many variables. It must be remembered that the data are being collected from secondary sources (clinical and pathology records, hospital discharge abstracts, death certificates) and NOT from the patients themselves. Thus, items of information that are not routinely available in these

Table 4.1. Basic information for cancer registries

Item	Comments
The person	
Personal identification[a]	
Name	According to local usage
Sex	
Date of birth or age	Estimate if not known
Demographic	
Address	Usual residence
Ethnic group[b]	When population consists of two or more groups
The tumour	
Incidence date	
Most valid basis of diagnosis	
Topography (site)	Primary tumour
Morphology (histology)	
Behaviour	
Source of information	For example, hospital record number, name of physician

[a] The minimum information collected is that which ensures that if the same individuals are reported again to the registry, they will be recognized as being the same person. This could also be a unique personal identification number.

[b] Ethnic group is included here because it is important for most registries, especially in developing countries.

Source: MacLennan (1991).

sources should be avoided. This applies especially to items of information that can be reliably recorded only by interviewing the patient (risk factors such as tobacco and alcohol use, diet, etc.), as well as those likely to be recorded in only a subset of cases (and not a random subset, at that), such as occupation or HIV status. As a general rule, unless reliable information can be collected on 80–90% of cases, the item should not be included in the registry data set. Some variables, although easy to capture, are of little relevance and are also best avoided (e.g. marital status). In *Cancer Registration: Principles and Methods*, a set of 10–11 essential variables is proposed (Table 4.1), and it is true that no cancer registry could function with less than this, so that these might be considered the ***minimum data set***.

However, a reasonable list of ***essential variables*** is more substantial than this. Table 4.2 is based on the recommendations of the European Network of Cancer Registries (http://www.encr.eu/images/docs/recommendations/recommendations.pdf).

There are many ***optional variables*** that might also be included, depending on specific local interests, bearing in mind considerations of the availability of the items of information in the data sources, as described above.

4. Coding

Several of the variables listed require coding, to facilitate analysis. For a number of the variables, standard, international coding schemes are available, and cancer registries should use them so that comparison of results between registries is possible.

The most important are the coding of the tumour (site, histology, behaviour, basis of diagnosis), using the International Classification of Diseases for Oncology (ICD-O), and the coding of stage, using the tumour–node–metastasis (TNM) system.

In addition, local coding schemes will be needed for:

- place of residence
- ethnic group (if recorded)
- source of information.

4.1 Classification of cancers – International Classification of Diseases for Oncology

Now in its third edition, ICD-O has been used for more than 35 years as the standard tool for coding diagnoses of neoplasms in cancer registries.

ICD-O is a multi-axial classification of the site, morphology, behaviour, and grading of neoplasms (and, in addition, it provides standard codes for the basis of diagnosis).

The topography code describes the **site of origin** of the neoplasm (the *primary site*, not the location of any metastasis) and uses the same three-character and four-character categories as in the neoplasm section of Chapter II of the International

Table 4.2. Essential variables for cancer registries

Item	Comments
The person	
Personal identification	In some countries a unique identification number, in others full name combined with date of birth and sex
Date of birth	Given as day, month, and year (dd/mm/yyyy)
Sex	Male (M) or female (F)
Ethnic group	According to local situation
Address including postal (or zip) code (and telephone number)	Needed for identification purposes and for geographically based studies
The tumour	
Incidence date	This date should be given priority, as outlined by the ENCR recommendations.
Primary tumour site	This should as a minimum be according to ICD-O.
Laterality	This should be recorded for all paired organs, but as a minimum for breast, eye, ovary, testis, and kidney (but observe the multiple primary rules).
Primary tumour histology	This should as a minimum be according to ICD-O.
Behaviour	This should as a minimum be according to ICD-O.
Basis of diagnosis	Most valid basis is recommended. All relevant methods may be recorded. The basis codes should be according to ICD-O.
Stage – (condensed TNM)	Stage is needed for international studies and for servicing clinicians. It is recommended to use the ENCR condensed TNM.
Initial therapy (i.e. initiated within 4 months from incidence date) [A clear manual on what is included should be available from the registry for all treatment items.]	As a minimum the registries should be able to present on a yes/no basis the treatment modalities used.
Surgery	Any surgical procedure of curative or palliative nature
Radiotherapy	Any radiotherapy of curative or palliative nature
Chemotherapy	Any cancer chemotherapy of curative or palliative nature
Endocrine (hormones)	Exogenous therapy, i.e. medication
Sources of information	
Sources of information	It is important to record ALL of the sources of information (hospital/institution) for each diagnosis and treatment modality in order to be able to do quality control, or to collect additional information. The relevant date and hospital/laboratory number are recorded for each.
Follow-up	
Last follow-up date	Needed to study follow-up (dd/mm/yyyy)
Vital status (at last follow-up date)	It may be of value to indicate whether known or assumed (e.g. based on linkages to death certificates) (dd/mm/yyyy)
Date of death	Needed to study survival and follow-up (dd/mm/yyyy)

ENCR, European Network of Cancer Registries; ICD-O, International Classification of Diseases for Oncology; TNM, tumour–node–metastasis.
Source: *Recommendations for a Standard Dataset for the European Network of Cancer Registries* (http://www.encr.eu/images/docs/recommendations/recommendations.pdf).

Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) classification of malignant neoplasms (except for those categories that relate to secondary neoplasms and to specified morphological types of tumours). ICD-O thus provides greater site detail for tumours than is provided in ICD-10. In contrast to ICD-10, ICD-O includes topography for sites of haematopoietic and reticuloendothelial tumours (as well as other cancers that, in ICD-10, are defined by histology, such as Kaposi sarcoma, melanoma, and sarcomas of soft tissue and bone).

The morphology axis provides five-digit codes ranging from M-8000/0 to M-9989/3. The first four digits indicate the specific histological term. The fifth digit, after the slash (/), is the behaviour code, which indicates whether a tumour is malignant, benign, in situ, or uncertain (whether benign or malignant).

A separate one-digit code is also provided for histological grading (differentiation).

The International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) book has five main sections. The first section provides

general instructions for using the coding systems and gives rules for their implementation in tumour registries and pathology laboratories. The second section includes the numerical list of topography codes, and the third section the numerical list of morphology codes. The combined alphabetical index provided in the fourth section gives codes for both topography and morphology and includes selected tumour-like lesions and conditions. The fifth section provides a guide to differences in morphology codes between the second and third editions of ICD-O.

To the greatest extent possible, ICD-O uses the nomenclature published in the World Health Organization Classification of Tumours series ("WHO Blue Books"). As these are revised, and new morphological terms introduced, new codes are prepared as updates/addenda to ICD-O, pending a fourth edition.

ICD-O has been published in a wide variety of languages (Chinese, Czech, English, Finnish, Flemish/Dutch, French, German, Japanese, Korean, Portuguese, Romanian, Spanish, and Turkish). It can be purchased from WHO (http://www.who.int/classifications/icd/adaptations/oncology/en/) or from the International Association of Cancer Registries, for members of that organization. A CSV file can be downloaded from the WHO website (http://apps.who.int/classifications/apps/icd/ClassificationDownloadNR/login.aspx?ReturnUrl=%2fclassifications%2fapps%2ficd%2fClassificationDownload%2fDLArea%2fDownload.aspx).

The IARC–IACR Cancer Registry Tools (IARCcrgTools) package includes *batch* programs for conversion from ICD-O to ICD-10. The conversion and check programs can only process text files having a *fixed field format*, although a File Transfer option allows conversion of a text file from delimited to fixed field format. The IARCcrgTools package is available from the website of IACR or IARC (http://www.iacr.com.fr/iacr_iarccrgtools.htm).

4.2 TNM coding system

The Union for International Cancer Control (UICC) TNM classification is the internationally accepted standard for cancer staging. It is an anatomically based system that records the primary and regional nodal extent of the tumour and the absence or presence of metastases.

Each individual aspect of TNM is termed a category:

- The T category describes the primary tumour site.
- The N category describes the regional lymph node involvement.
- The M category describes the presence or otherwise of distant metastatic spread.

Cancer staging is important not only for clinical practice; it also provides vital information for policy-makers developing or implementing cancer control and prevention plans, and it is therefore important to include the TNM classification as part of cancer registration.

The TNM classification is regularly updated and is now in its seventh revision.

The UICC website gives an explanation of the use of the TNM classification system and how to obtain the relevant coding manuals (http://www.uicc.org/resources/tnm).

Cancer registrars in LMICs may have difficulty in abstracting the full TNM code from clinical records, if this has not been explicitly recorded by the clinicians or pathologists. For this reason, a simplified version has been created by the European Network of Cancer Registries: the condensed TNM, available in English and French (http://www.encr.eu/images/docs/recommendations/extentofdisease.pdf).

This version allows for recording of T and/or N and/or M when they have not been explicitly recorded in the clinical or pathological records. The cancer registry then attempts to score extent of disease according to the condensed TNM scheme:

T:	**L**	**A**	**X**
N:	**0**	**+**	**X**
M:	**0**	**+**	**X**

(**A**, advanced; **L**, localized; **X**, cannot be assessed), where T and N are extracted, if possible, from the pathology report, or, in its absence, from the clinical record (endoscopy, X-ray, etc.). M is based on the best available information, whether clinical, instrumental, or pathological. For M, clinical signs and findings are enough to justify M**+** in the absence of pathological confirmation of metastatic deposits.

Both the full TNM and the condensed TNM allow tumour extent to be expressed according to the familiar numerical staging scheme:

I Tumour localized (T**L**/N**0**/M**0**)
II Tumour with local spread (T**A**/N**0**/M**0**)
III Tumour with regional spread (any T/N**+**/M**0**)
IV Advanced cancer (metastatic) (any T/any N/M**+**).

4.3 Local coding schemes

4.3.1 Place of residence

"Place of residence" codes should correspond to national subdivisions of the population as they appear in national statistical publications and for which there is information available on the size and composition of the population. It may be possible to develop a hierarchical coding scheme, where there are several levels of population subdivision (region, province, district, city ward, etc.).

4.3.2 Ethnic group

"Ethnic group" codes should correspond, if possible, to any official

categories recognized in national statistical publications, especially if there is information available on the size and composition of the population, by ethnic group.

4.3.3 Source of information

The codes for "source of information" will almost always be specific to the cancer registry, and will have to be developed by the registry itself. Careful thought should be given to developing a hierarchical coding scheme that will facilitate extraction of information (e.g. lists of cases) from the registry database, and tracing the records of lists of cases.

Thus, a coding scheme might aim to have different levels, such as:

1. Type of source (hospital, diagnostic laboratory, death certificate)
1.1. List of hospitals – public
1.2. List of hospitals – private
1.3. Hospices
1.1.1. Clinical services (medicine, surgery, radiotherapy, etc.).

It is important, when developing the coding scheme, to allow for its expansion in the future, as new sources of case data are included, while respecting the structure of the coding scheme.

As noted above (Table 4.2), the registry will record the case record number, but unless it is clear to which hospital/service or laboratory this number refers, it will be very difficult to trace the record if it is required for extracting additional information, for correcting errors in the registry database, or for research purposes.

5. Information on the population at risk

As described in Chapter 3, the registry should maintain a population file, which, for each calendar year, contains the population estimate for every combination of:

- ethnic group (if applicable)
- sex
- age (standard 5-year age groups, IF possible, separating infants [age, 0 years] from children [age, 1–4 years]) and including the numbers of persons of unknown age.

Key points

- A key feature of the PBCR is the use of multiple sources of information on cancer cases in the target population. Registry procedures allow identification of the same cancer case from different sources (while avoiding duplicate registrations). The sources can be grouped into three broad categories: hospitals, laboratories, and death certificates.
- Most registries use a mixture of active and passive methods of case finding.
- The development of computerized health information systems may provide some scope to use electronic databases for case finding.
- Cancer registries set out to record data for a set of variables on each cancer case. There is a uniform tendency, when a cancer registry is planned, to aim for too many variables.
- There are some 17–20 variables that it is essential for a registry to collect on each case registered. Additional, "optional" variables should be kept to a minimum. Several of the variables listed require coding, to facilitate analysis. Standard, international coding schemes are available for some variables, and cancer registries should use them so that comparison of results between registries is possible.
- The most important are the coding of the tumour (site, histology, behaviour, basis of diagnosis), using ICD-O, and the coding of stage, using the TNM system.