# CHAPTER 5
# Managing and controlling the database

## Contents

# CHAPTER 5
# Managing and controlling the database

## Population base

Correct sizing of denominators, specifically with regard to diagnostic and care programs aimed at the entire population present (screening), is integral to the epidemiological controls required from the registries. Furthermore, the emerging need for special surveys in at-risk population groups (the elderly, immigrants, socially vulnerable groups) placing increasing pressure on registries to ensure that they can provide this crucial service.

Considering the complexity (discussed earlier) of the estimates required to quantify the population effectively cared for and present in the area as compared with traditional information based on registered residency information, it is helpful for each registry to link up with the institutions that deal with demographic assessments in the area (towns, regions).

Customarily, the population base of each cancer registry is the population officially residing inside its catchment area for the years covered by its records. In accordance with the homogeneity criteria for the national incidence database, the data available must include:

- sex: 1 = male; 2 = female;
- year of residence: four-digit format;
- age class: preferably annual (see below);
- town of residence: ISTAT code *(*use the reference file in the Appendix);
- number of subjects: by sex, year, age, town of residence.

The age classes should be disaggregated by individual year. Alternatively, ages may be grouped into classes of up to five years, preferably providing individual data for the 0-1 year class (e.g., 0; 1-4; 5-9; 10-14, etc.); in this case the five-year classes must extend at least to the "85 years and beyond" bracket, and must of course indicate any deviations from this format.

### Data sources

The populations must be collected from official sources, according to the following hierarchical classification:

- ISTAT;
- regional source;
- municipal source;
- other sources (e.g., local estimates).

The data source used must be specified in the methods presentation section.

### Resident and present population

Reference to the population officially residing inside the area covered by the registry has always been a guideline for the quality of the data available. Having civil status information organized by the municipal administrations guarantees the quality of the information, in the sense of its alignment with reality.

As is well known, in addition to residents, the national health service also provides care to other "mobile" population segments, including people domiciled inside the area and enrolled with the registry office of the regional USL Agencies (Local Health Units). In recent years, this segment of the population has grown progressively due to the immigration of individuals (and families) who, while not possessing the requirements for recognition of citizenship, are in fact a stable part of the area in which they receive health care.

This population is interesting from the point of view of oncological epidemiology, since it often has different risk levels than the native population. In addition, with respect to the principle of fairness in the right to health, these social groups are also being targeted for prevention efforts.

Consequently, the progressive spread of oncological screening through the country results in considerations of the disease in social strata that are stably integrated in an area and often with greater individual risks.

Recording of these cases must therefore line up with the possibility of identifying the reference population (**present** and with access to care, in addition to being **resident**) with the lowest level of distortion.

The towns, USL agencies, provinces, and regions generally provide data that will become increasingly necessary for the registries, but at the moment acquisition of these data is often only in the planning stages. This both justifies efforts by and encourages commitment from individual entities to monitor the course of demographic indicators locally.

## Indicators

In detail, the indices used concern:

### Number of cases

Indicates the total number of cases recorded:

$$N = \sum_i n_i$$

($n_i$ =number of cases per age class; $i$=index of the five-year age class)

### Proportional distribution

Indicates the site-specific percentage level of incident cases and deaths compared with total cases recorded.

### Crude rate

Rate per 100,000 inhabitants per year:

$$T = \frac{\sum_i n_i}{\sum_i p_i} \times 100,000$$

($p_i$= population by age class)

### Standardized rate

Rate per 100,000 inhabitants per year, standardized by age using the direct method (for a reference population). This makes it possible to compare different territories (referring to the same standard population), eliminating the effect of different age compositions of the populations:

$$T_{st} = \frac{\sum_i (T_i \times \text{standard pop. } i)}{\sum_i \text{standard pop. } i}$$

Any standard population can be selected for comparison with the data available; it is however advisable to refer to the global, European, or Italian population model for the best comparison of the data.

### Cumulative risk

Expresses the probability of onset of cancer between birth and a specific age (likelihood of becoming ill if death does not occur due to other causes). This is often expressed as risk between 0 and 74 years, per 1,000 inhabitants:

$$R_{cum} = 1 - \exp\left[\left(-\sum_i T_i\right) \times 5\right]$$

### Mortality/incidence ratio

Expresses the ratio between deceased and incident cases (generally expressed by type and site).

### Years of life lost/gained

Estimate of the years of life lost or gained by a cohort exposed to the determinant compared to a non-exposed cohort. Years of life lost (YLL) or potentially lost compared with an expectation (YPLL). Working from this estimate we can calculate the average years of life lost/gained (average of YLL/YPLL of a population), the crude rates of YLL, YPLL (relating YLL/YPLL to the population younger than the selected limit), the standard rates (direct standardization of YLL/YPLL), or the cumulative rates of YLL/YPLL.

### Survival

At set intervals it is helpful to verify survivorship in the recorded incident cases in order to conduct internal controls on the data and for epidemiological analysis.
Patient follow-up must be conducted through a certified archive (town of residence, assisting USL Agency if linked to the towns). The follow-up data in the case record format must be supplemented with a variable indicating the patient's life status (alive, deceased, lost, etc.). The final summarized value can be expressed as **observed** survival (according to the actuarial model or Kaplan-Meier) or relative survival; the latter is expressed as the ratio between **observed** and **expected** survival. Expected survival is calculated using the general mortality rates. Relative survival therefore represents survival of the cohort of patients in question net of mortality due to other causes.
The algorithms of these indexes and the corresponding standard errors are generally indicated by the various calculation programs available.
Period survival is an alternate approach to analyzing a cohort of incident patients (methodological details are provided in the SEERStat program).
Cases recorded based on the date of death (DCO, autopsy cases) are excluded from the survival analysis.

### Prevalence

Represents (proportionally or as an absolute number) the number of patients with previous diagnosis of cancer (within a number of years to be determined) still alive at the time of observation. This is an important index for planning care. For patients lost at follow-up or considered censored for other reasons, the estimates generally used are based on survival data. A number of methods are also available for calculating indices for patients with multiple cancers.

### Time trends

The precise indicators that can be calculated using the registry archive are considerably more informative if they are placed in a temporal context. Thus, in recent years the analysis of time trends (incidence, mortality, survival, prevalence) has provided increasingly valuable elements for interpreting the incidence of cancers in various geographic contexts. An initial impression can thus be drawn from tables or graphics showing the rates (crude or standardized) for the periods as they are measured.

Causal fluctuations in the indices, especially for restricted conditions with large confidence intervals, often do not provide convincing information. Therefore, to get a concise picture of progress over time, it is helpful to use a regression model. Among the many possible solutions, one way to standardize the procedure is offered by the *Joinpoint Regression Program*,[1] which is based on identifying the linear segments that adapt best to the rates observed (rates logarithm), minimizing the sum of the squares of the distances of the points from the segments.[2] The trend can be broken down into a maximum of *k* segments in *joinpoint*, set in advance. The *joinpoint* represents the year that identifies a change in the trend.
The model can be represented in a single equation:

$$\ln(T_{stand}) = \beta_0 + \beta_1 x_i + \delta_1 (x_i - \tau_1)^+ + \\ + \delta_2 (x_i - \tau_2)^+ + \varepsilon_i^{(k)}$$

For further information on this method, please refer to the bibliography (or go to srab.cancer.gov/joinpoint).

### Standard errors and confidence intervals for the indicators

The programs used to calculate indicators generally provide for calculation of standard errors (SE) and confidence intervals, referring to 95% probability. This information is very useful for all indicators and if possible should not be omitted in order to explain the causal fluctuation of the precise indicator (rate, probability, etc.), especially when a small number of cases are available (typical of many regional situations) for better assessment of any actual differences in the indicators.

#### Crude rate SE

The calculation assumes the distribution of cases according to the Poisson model:

$$SE_{crude} = N / population \times 100{,}000$$

#### Standardized rate SE

The SE is calculated assuming that the number of cases has a Poisson probability distribution. Supposing that the rate adjusted by age is composed of age groups ranging from x to y:

$$SE_{SR_{x-y}} = \left[ \sum_{i=x}^{y} \left( \frac{w_i}{\sum_{j=x}^{y} w_j} \right)^2 \times \left( \frac{N_i}{\text{population}_i^2} \right) \right]^{\frac{1}{2}} \times \\ \times 100{,}000$$

#### Cumulative rate SE

$$SE = 5 \times \sqrt{\sum_i \frac{n_i}{P_i^2}}$$

## Index calculation programs

These are a significant resource, because they have a positive effect on the time and costs needed to produce registry reports. With regard to the many solutions offered by software programs, both commercial and non, a few should be mentioned here.

### *CanReg*

This program is designed for global management of the registry, and is available free of charge from the IARC/IACR. It includes an analysis function with frequency distribution, incidence tables, and an interface with EpiInfo® for the most complex processing procedures.[3]

### *SEERStat*

This is currently the most complete analysis software for the cancer registry, available free of charge from the manufacturer's website.[4] It requires data to be loaded using a special program (*SEERPrep*).[5] Its crucial features are:

- high qualified and skillful manufacturers;
- designed specifically for the needs of a cancer registry;
- provides all of the usual statistics and options for complex analyses;
- easy to use;
- free;
- updates are released periodically;
- can be used with any personal computer.

The software uses a database in which the variables can be converted both during and after loading, and it includes analysis modules for:

- frequency distribution;
- crude and standardized rates;
- survival analysis, observed and relative, by cohort and by period;
- duration prevalence, with the option to assess patients with multiple tumors.

A utility is also available for exporting data and tables to other software (e.g., Office) for further analysis or to create graphics.
SEER also offers other data analysis packages that work directly from tables produced by *SEERStat*:

- *DevCan*: probability of developing cancer or of dying from it;
- *Joinpoint*: analysis of temporal trends;
- *CanSurv*: Survival analysis with graphics and models (Standard parametric, Cox, Mixture cure);

- *ComPrev*: complete prevalence;
- *ProjPrev*: *SEERStat* duration prevalence projections on other populations.

The programs come with extensive documentation.
Additional documentation, including examples and the theory behind the models, is available at www.seer.cancer.gov.

# Error control

## Data entry procedures

The different filing systems used in registries presuppose different error control criteria when entering data. Compared to traditional ways of identifying cases (through systematic review of medical records and subsequent entry of individual cases into electronic media), it is increasingly common for registries to acquire entire sections of other healthcare databases (HDD, anatomical pathology reports, death records), in which data can be more easily controlled after the fact using automatic programs that will be described below.
In order to prevent procedural errors during the many steps in acquiring the data and managing the database, it also appears to be very important that each registry formalize all phases of its work in a handbook of procedures, which on the one hand can offer a guarantee of reproducibility and consistent quality over time in the data processing system, and on the other hand can stimulate more active involvement of personnel and skill update training in all work phases.

## Programs for controlling the precision of data

The IARC periodically issues updated control programs that can detect most of the possible consistency errors within individual records. These programs are freely available to the registries and can be downloaded directly from the Internet.
Independently of the utility of using these programs when initially entering series of data obtained from external sources, the registries must control their own incidence data **before** sending them to the national database or other research projects (e.g., Cancer incidence, EUROCARE).
The inset start on this page lists the leading software programs available today that guarantee complementary data control procedures (thus the use of both is required), in addition to the option for transcoding between the various classification systems (various versions of ICD and ICD-O) and control over multiple tumors.

### Data Control Software

***IARCcrgTools****

Software compatible with Windows versions 95/98/Me (preferably NT/2000/XP), which includes:

- a program for converting among the classifications:
  - from ICD-9 (1975) to ICD-O-2 (1990)
  - from ICD-9 (1975) and ICD-O-1 morphology (1976) to ICD-O-2 (1990)
  - from ICD-10 (1992) to ICD-O-2 (1990)
  - from ICD-10 (1992) and ICD-O-2 morphology (1990) to ICD-O-2 (1990)
  - from ICD-O-1 (1976) to ICD-O-2 (1990)
  - from ICD-O field trial edition (1988) to ICD-O-2 (1990)
  - from ICD-O-2 (1990) to ICD-9 (1975)
  - from ICD-O-2 (1990) to ICD-10 (1992)
  - from ICD-O-2 (1990) to ICD-O-3 (2000)
  - from ICD-O-3 (2000) to ICD-10 (1992)
- a program for checking validity and consistency among the variables:
  - age, incidence, and date of birth
  - age, anatomic site, and morphology (ICD-O-3)
  - sex and anatomic site
  - sex and morphology (ICD-O-3)
  - behavior and anatomic site (ICD-O-3)
  - behavior and morphology (ICD-O-3)
  - grade and morphology (ICD-O 3)
  - basis of diagnosis and morphology (ICD-O 3)
- a program for monitoring multiple tumors, in accordance with the 2004 IARC/IACR rules.[6]

The program is easy to run and interpret and features a good level of internal documentation, which is also available on www.iacr.com.fr/ .[7]

**version 2.03, January 2006*

***DEPedits****

Similar to the above, it includes programs for converting between the various versions of ICD and ICD-O and transcoding of these to ICD-O-3 (with limitations with respect to the new entities introduced by the latter, concerning in particular the lymphohematopoietic system).[8]

It offers a control program based on ICD-O-3, which is required before data can be sent to the IARC Descriptive Epidemiology Group (DEP). It highlights both errors in the attribution of codes for topography, morphology, and sex, as well as (in a separate format) cases with "unusual" combinations tagged for additional checks.

This software offers all the conversions and controls available in IARCcrgTools, as well as:

- special controls on survival variables;
- special controls on juvenile cancers;
- validation (optional) according to the ICD-O-3 topographic and morphological combination criteria consistent with the validation list adopted by SEER (Surveillance Epidemiology and End Results, National Cancer Institute, USA).[9] The criteria required for this operation are specified in the documentation provided with the program and on the SEER website.

Specifically, DEPedits checks the following:

- validity:
  - sex
  - incidence date
  - date of birth
  - date of follow-up
  - ICD-O-3 topography
  - ICD-O-3 morphology (first 4 morphology digits)
  - ICD-O-3 behavior (5th morphology digit)
- consistency:
  - sex ↔ICD-O-3 T
  - sex ↔ICD-O-3 M
  - ICD-O-3 T ↔ICD-O-3 M (IARC)
  - ICD-O-3 T ↔ICD-O-3 M (SEER)
  - ICD-O-3 M ↔basis of diagnosis
  - specific morphological diagnoses generally allowed if histocytological confirmation level is provided, with the exception of the brain and liver, which can be accepted even with clinical diagnostics
  - date of incidence ≥ date of birth
  - date of incidence ≤ follow-up date
  - age ↔ICD-O-3 T
  - age ↔ICD-O-3 M
  - life status ↔basis of diagnosis

This program is also fairly easy to use and features an extensive explanation of its procedures and complete instructions; it is available at www.ENCR.com.fr/download.htm

** version 1.00, February 2006*

***CHILD CHECK Program***

This program is available directly from the IARC. It focuses on juvenile tumors and is designed to monitor the consistency of individual records and convert data from ICD-O to the *International Classification of Childhood Cancer* (ICCC).[10]

The control procedures check:

- the codes defined for dates, age, and sex
- the accuracy of morphology and topography codes
- unlikely combinations of cancer type, sex, and age
- unusual topography codes for specific morphologies
- inappropriate use of non-specific morphology codes

The conversion procedures convert codes:

- from ICD-O-1 (1976) to ICD-O-2 (1990)
- from ICD-O-2 (1990) to ICCC (1996).

**DATA CONSISTENCY CONTROL PROGRAMS**

***CheckRT***: AIRTUM-CCM Software for quality control
(Ivan Rashid, Modena Cancer Registry)
*CheckRT* is a program that allows cancer registry staff to quickly run in-depth quality controls on data.
The purpose of *CheckRT* is to formalize, publish, and expand the criteria used to accredit Italian cancer registries with the Italian Association of Cancer Registries (AIRTUM), and providing registries that are already accredited with a self-assessment tool.
*CheckRT* can import incident case history, mortality, and population information in Microsoft Access format and generates detailed reports in text format about the checks conducted on the data.
Checks conducted
CheckRT *version 4.0 can run the following checks on registry data:*

- analysis of the topography subsite (for a total of 191 checks)
- analysis of the percentage of generic and poorly defined sites (82)
- analysis of the DCO percentage (146)
- analysis of the microscopic confirmation percentage (150)
- analysis of the percentage of specific morphologies by site (95)
- analysis of the percentage of generic morphologies (46)
- analysis of the value of the mortality/incidence ratio (58)
- analysis of the trend in the mortality/incidence ratio (58)
- analysis of the stability of the mortality/incidence ratio (58)
- analysis of the standardized incidence rate by site (73)
- analysis of the male/female ratio (38)
- analysis of the distribution of the age-specific rate (72)
- analysis of the stability of microscopic confirmations (72)
- analysis of childhood cancers (19)
- analysis of survival a year after diagnosis (72)
- analysis of the stability of the standardized incidence rate (73)
- analysis of the EUROCARE site-morphology-control inconsistencies (57)

for a total of 1,360 checks. Each individual check is associated with an approximate score that indicates its weight in the controls group. The sum of these points, called the benchmark, is a concise indicator of the quality of the cancer registry.
Refer to the Appendix to consult the user manual.

## Checking the completeness of records

The premise behind a cancer registry for a population and its degree of usefulness lie in full coverage of registration in the registry's catchment area. This methodological premise requires that an assessment be conducted after recording and control procedures are complete (and before data is published) to remove any selection bias from the afferent case histories.
In addition to checks on the accuracy and internal consistency of the processed data, the scientific community of cancer registries has long used quality indicators for effective and efficient flow of information. When data are produced by the registries, they must be accompanied by these indicators, which indicate the reliability of the procedures used to acquire and check incident cases. The qualitative assessment of registry data (high level of specificity, in observance of the rules of registration) implies a verification of the completeness of the record (high level of sensitivity in intercepting incident cases).
A series of variables, most of which have been addressed in previous chapters and which are among the frequency indicators, can be extrapolated critically with the goal of achieving compliance with shared registration rules (reproducibility, objectivity) and ensuring a high level of informative detail along with the highest level of completeness (in intercepting all actual incident cases) and coverage of its target population.[12] This type of control must be part of the standard procedures of a registry and not just take place on the occasion of congresses or other scientific events.

### Completeness of coverage

The primary purpose of this control is to identify any missed cases (but also accidental duplication of cases or the inclusion of ineligible cases) caused by problems related to the flow of information. Naturally, the quality and consistency control procedures listed above can be used for this purpose, as well as a number of different approaches specifically dedicated to this purpose.

#### Proportion of cases with microscopic confirmation

This indicates the quality of the incidence registered, which rises with the percentage of cases that reach the golden standard of a histopathological and cyto-pathological diagnosis. Hematological diagnoses obtained through bone marrow puncture (cytology) or bone marrow biopsy (histology) should be assimilated at this level.
In relatively recent years some diagnostic techniques (CT, MRI, markers) have become widespread that in effect achieve the level of microscopic quality, making

it possible to identify the lesion and consequently to plan treatment. In particular, liver and brain cancer are the sectors in which these methods can ensure diagnostic quality similar to histopathology: this level of diagnosis makes it possible to provide a specific morphology code for the tumor that is otherwise admissible only with microscopic confirmation. The percentage of cases with microscopic diagnosis (total and site-specific) must demonstrate substantial homogeneity among bordering geographical areas, short of clear causes; it may be lower in cases where the registry has not fully consulted the cyto-histopathological documentation, whereas an excessively high percentage can mask under-registration due to loss of cases with a lower diagnostic level.
A careful examination of this indicator can point to the checks and corrective measures needed to adjust the flow of information starting from the earliest phases in which incidence information is generated.

**DATA LINKAGE PROGRAMS**

*Software for Automated Linkage in Italy SALI*
(Dr. Luigino Dal Maso CRO Aviano)
Procedures for linking cases from different archives offer an especially helpful opportunity to share information provided by different sources, which can be used for epidemiological studies and surveillance of patients recorded in disease records. Where unambiguous codes are used (deterministic criterion) these procedures can be handled by most records-management software programs. In the vastly more common event that reliable key fields are unavailable, it becomes necessary to use software that can process a medium-high number of records using common personal identification data, even in cases where they do not correspond perfectly, ensuring minimal loss of possible linkages.

Developed by the epidemiology and biostatistical unit at the Oncology Reference Center in Aviano, the *SALI* program was created in order to link individual records in average size registries (on the order of 100,000 records), allowing for manual revision of results and safeguarding, in every stage of the operation, the personal confidentiality of the individuals whose data is being processed.[11]
The program is optimized for a linkage probability of less than 1%, although it is able to provide good results even with higher expected percentages (subject to the power of the processor).
Developed in CA-Clipper language, SALI uses database-format records and requires surname, name and date of birth as key fields, making it possible to consider possible attribution errors in the key fields.

The linkage procedure is based on seven levels, of which two are automatic and five are interactive, in which the user is presented with a number of windows in which to decide whether to accept or reject the proposed link.
*SALI* can be used in any IBM-compatible operating system (DOS or Windows).
Refer to the Appendix to consult the user manual.

### DCI/DCO cases

The integration of various sources of information used by a registry generally leads to a number of incident cases collected solely from the patient's death certificate. These cases, with the exception of errors when certifying the death, were not recorded during life (through clinical examinations, hospitalizations, etc.), and generally indicate problems in the flow of information.
Returning to the previous discussion, an initial check generates a class of cases called DCI cases (or DCN: *death certification initiated/notification*), on which trace-back controls must be done to recover any documentation that was missed in the patient's diagnostic-care path. After this step, in which a portion of the DCI cases are reintegrated into the incidence with a better diagnostic level, a number of deaths from cancer remain called DCO (*death certification only*) cases, for which checks did not locate any indication of diagnosis during life.
DCI and DCO are important indicators of the quality of the flow of information in the registry, and these levels must be recorded separately. DCI cases indicate a failure in the registry with regard to completeness of detection; DCO cases also highlight possible deficiencies inherent to the health information systems consulted. The final inclusion of DCI/DCO cases in incidence, however, only partially corrects the completeness defect, which most likely includes a percentage of cases lost because they are not deceased. By way of example, granting a percentage of 10 lost in 100 and mortality of 60%, six of these cases will be subsequently recovered as DCI/DCO, while four will remain unknown, putting completeness of detection at 96%.
The accuracy of death certificates, while generally inferior to incidence data, can similarly be tracked by the registry using a cross comparison of its own information coming from clinical sources.

### Mortality/incidence ratio (M/I)

Because of its stability over time and in different geographical areas, this index is one of the leading and most reliable assessment standards. The reference to information in the literature allows registries (especially during the start-up phase) to quickly identify the loss of incident cases (with an increase in the M/I ratio) or the

accidental inclusion of duplicates and prevalent cases, or, alternatively, incomplete availability of deceased cases (low M/I). It is helpful to stress that it is absolutely necessary, for all checks working from mortality data, that these data be certified and come from a source outside the registry (ISTAT, region, local health units, etc.).

#### Case entry date

This is an important variable for controlling the completeness of information in relation to publication dates. In fact, traditional five-year scanning in the production of information by registries, often with significant delays between the onset of cases and their recording, has been progressively abandoned in recent years in favor of faster data production, often prompted by the assessment needs of screening programs or requirements to adapt health planning more quickly as a function of certain epidemiological emergencies. As data production times shorten, the risk of an incomplete reconstruction of the full record of cases increases, with obvious problems in terms of completeness that translate into an increase in cases "retrieved" after the incidence is closed.

A need to dynamically control completeness of incidence data as a function of time and of production times is addressed by one of the essential variables, which indicates the **case recording date** (entered into the record). This makes it possible to develop a function (with percentage of cases recorded on the Y axis [ordinate] and time on the X axis [abscissa]) that beginning at Time T0 (when incidence began to be recorded) indicates the rate at which a particular level of completeness was achieved (represented by the upper plateau of the curve in the stabilization phase of the entry of new cases). This makes it possible to continuously monitor the rate at which completeness is achieved (and, consequently, the minimum interval that can elapse between incidence and its recording), efficiently representing the reliability of the information in terms of accuracy and punctuality.

#### Incidence by registration sub-areas

This makes it possible to detect differences in case recruitment consistency in the catchment area (difference between various agencies, towns, provinces, etc.).

#### Sample checks

These are conducted by selecting a sample of oncology patients from the case histories in an area (hospitals, clinics, healthcare flows) and checking that they were correctly reported by the registry.

#### Survival

When registries run this analysis regularly, among other things, it allows them to identify possible completeness or recruitment problems in the incident cases by monitoring discrepancies from expected values.

#### Temporal incidence trends

The study of temporal incidence trends is a useful tool for checking any qualitative or quantitative changes over time in the flow of information belonging to the registry. In the past, unexpected temporal incidence changes were generally indicative of under- or over-registration problems; in recent years the introduction and spread of more sensitive diagnostic techniques (endoscopy, ultrasonography, tumor markers, radiology) or the implementation of screening programs (spontaneous or organized) in the population have led to rapid and considerable increases in the incidence of certain forms of tumors, perfectly in line with expectations based on previous experience.

Every temporal incidence change (general and site-specific) must in any case be considered and interpreted carefully and thoroughly. The recent availability of automatic programs to aid this kind of control can only help improve and noticeably standardize quality assessment procedures, both for emerging registries and for those with well-established histories.

#### Special techniques

Capture/recapture *models* or other statistical techniques will be explained below.

### Completeness of detail

The negative effect of missing information on recording accuracy is on a par with the negative effect of errors. It may not in fact always be possible to complete all the information the archive requests for each individual patient. Certain variables can therefore be "missing" from cases, sometimes representing a serious deficit (vital information such as sex, residence, date of birth, or date of incidence), and sometimes less serious (supplementary information).

Every registry must pay attention to the frequency of missing information, because it is an important part of registration quality control.

In terms of monitoring missing values of variables handled by the registry, an important section concerns two crucial variables in particular, which generally fall under international quality control protocols: the percentage of cases recorded with an unknown or ill-defined site and the percentage of patients of unknown age.

The first group of cases, indices of the quality of diagnostic information, specifically regard codes ICD-10 and ICD-O C26, C39, C76, C77, and C80.

The percentage of patients with an unknown age at diagnosis (because date of birth is missing) generally points to a serious lack of identifying information about

the patient. For this reason, in registries in developed countries consensus is that this index must be kept well below 1%.

Finally, there are other missing items for which the percentage must be carefully monitored, and in particular:

- sex;
- town/province of birth (migrant control);
- tumor histotype (percentage of neoplasms with generic "NOS" code);
- residence;
- incidence date.

Increasingly close collaboration between registries and the territorial health organization has also led over time to a need to produce detailed information about additional variables that are useful in assessing the impact of diagnosis (e.g. screening) or other healthcare procedures:

- additional topographical and morphological details;
- additional disaggregations of staging variables (size, number of lymph nodes, distant metastasis sites, etc.);
- screening status;
- treatments;
- socio-economic variables;

for which routine monitoring processes should be duly organized.

Registries must periodically check the completeness of their registration using shared quality indicators and the control systems described in this handbook. Clearly, the central AIRTUM database can also integrate these activities through other ad hoc control systems, including that of the North American Association of Central Cancer Registries (NAACCR).[13]

## Revisions and Updates

### Follow-up

Updates on the life status of oncology patients has long been a routine part of registry procedures, both for quality control of the information, and to produce survival data. The registry can run this update at regular intervals (generally every two years), indicating the date of the last report on the patient and his/her life status at that time.

Thus, using only two variables, all three possible solutions can be represented:

- patient alive at the conclusion of follow-up (follow-up closure date with patient alive);
- death of patient (date of death with life status indicating the occurrence of death);
- censoring of the patient before closure of follow-up (date of last report about the patient with patient reported as alive); this event occurs largely in two situations:
  - death of the patient due to causes not related to the cancer;
  - loss of patient to follow-up for any reason;

in this case therefore, all effective survival reached by the patient up through the most recent available date is considered, and he/she is not considered in any case to be deceased due to the cause under examination (cancer).

Obviously, the variable indicating life status can be organized differently in order to express, for example, the percentage of patients who are deceased, emigrated, disappeared, etc., yet still preserve a single event mode in terms of the cause under examination (death from tumor). The "cause of death" variable can function as a substitute in this sense, because it can guarantee (depending on the reliability of the information) the construction of cause-specific survival not considering the effect of competitive mortality.

It is clear that the source of the life status data must be completely reliable and must generally be the same source that determines the patient's other demographic variables (municipal registers or files directly derived from them), while in the case of follow-up aimed at other end points (recurrence of disease, onset of complications), the source of the data must be explicitly stated (clinical units, health information systems) and must be subjected to completeness control procedures analogous to those conducted for the other current information in the registry.

## References


1. Statistical research and application branch, Division of Cancer control and population sciences, National Cancer Institute, USA (srab.cancer.gov/joinpoint/).
2. Lerman PM. Fitting Segmented Regression Models by Grid Search. *Applied Statistics* 1980; 39: 77-84.
3. Cooke AP, Parkin DM, Ferlay J. CanReg 4, Descriptive Epidemiology Unit, IARC/IACR. Lyon 2005. (www.iacr.com.fr/canreg4.htm).
4. Surveillance research program. National Cancer Institute SEER*Stat software (www.seer.cancer.gov/seerstat) version 6.1.4, 2005.
5. Information Management Services Inc. The SEER Program. National Cancer Institute. v.2.3.2, 2005.
6. IARC/IACR/ENCR Working group. *International Rules for Multiple Primary Cancers (ICD-O third edition).* IARC Internal report 2004/02, Lyon 2004.
7. Ferlay J, Burkhard C, Whelan S, Parkin DM. *Check and Conversion Programs for Cancer Registries (IARC/IACR Tools for cancer registries).* IARC Technical report no. 42, Lyon 2005.
8. Ferlay J. DEPedits 1.00, IARC, Lyon 2006.
9. Surveillance Epidemiology and End Results, National Cancer Institute, USA (seer.cancer.gov).
10. Kramarova E, Stiller CA, Ferlay J et al. *International Classification of Childhood Cancer.* IARC Technical report no. 29, IARC Lyon 1996 (disk included).
11. Dal Maso L, Braga C, Franceschi S. Methodology Used for 'Software for Automated Linkage in Italy' (SALI). *Journal of Biomedical Informatics* 2001; 34: 387-95.
12. Jensen OM, Parkin DM, MacLennan R et al. *Cancer Registration, Principles and Methods.* IARC Scientific publication no. 95, Lyon 1991.
13. Havener LA. Standards for Cancer Registries Vol. III. *Standards for Completeness, Quality, Analysis and Management of Data.* North American Association of Cancer Registries Inc, 2004