EUROPEAN JOURNAL OF CANCER 45 (2009) 747–755

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Review

# Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness

*Freddie Bray*[a,b,*], *D. Max Parkin*[c,d]

[a]*Department of Clinical and Registry-based Research, Cancer Registry of Norway, Institute of Population-based Cancer Research, Montebello, N-310 Oslo, Norway*
[b]*Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway*
[c]*Clinical Trials Service Unit and Epidemiological Studies Unit, University of Oxford, Oxford, UK*
[d]*Cancer Research UK, Centre for Epidemiology, Mathematics and Statistics, London, UK*

ARTICLE INFO

*Article history:*
Received 1 September 2008
Received in revised form
5 November 2008
Accepted 10 November 2008
Available online 29 December 2008

*Keywords:*
Cancer registry
Data quality
Comparability
Completeness
Validity
Timeliness
Quality control

ABSTRACT

The value of the modern cancer registry and its ability to carry out cancer control activities rely heavily on the underlying quality of its data and the quality control procedures in place. This two-part review provides an update of the practical aspects and techniques for addressing data quality at the cancer registry. This first installment of the review examines the factors influencing three of the four key aspects, namely, the comparability, validity and timeliness of cancer registry data. Comparability of cancer data may be established through a comprehensive review of the registration routines in place. Validity is examined via numerical indices of that permit comparisons with other registries, or, within a registry, over time, or with respect to specified subsets of cases. There are no international guidelines for timeliness at present, although specific standards for the abstraction and reporting of registry have been set out by certain organisations.



## 1. Introduction

The role of the population-based cancer registry has evolved beyond that of a provider of information on cancer incidence within a defined catchment area.[1] Given sufficient resources, the modern registry is active in a number of areas of cancer control, including epidemiological research on the causes of cancer, the monitoring and evaluation of screening programmes, and the follow-up of cancer patients in relation to the quality of cancer care they receive.[2,3] The value of a cancer registry and its ability to carry out such activities rely heavily on the quality of the data, and the quality control procedures in place.[4,5] Registries with a wide portfolio of activities will tend to improve the quality of their routine statistics simply by utilising the collected data, as well as by activating interest amongst collaborators (clinicians and researchers) in the daily registration procedures.[6]

As Skeet[7] writes in the standard text on registry methodology, 'all registries should be able to quote some objective measure of (ascertainment) rather than relying on received

* *Corresponding author:* Address: Department of Clinical and Registry-based Research, Cancer Registry of Norway, Institute of Population-based Cancer Research, Montebello, N-310 Oslo, Norway.
E-mail address: freddie.bray@kreftregisteret.no (F. Bray).
0959-8049/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2008.11.032

wisdom and pious hope'. This advice is not always heeded, however, and we take the opportunity here to provide some guidance on the practical aspects and techniques for addressing data quality at the cancer registry. This two-part review serves as a condensed update of the International Agency for Research on Cancer (IARC) Technical Report on the subject 'Comparability and Quality Control in Cancer Registration' published in 1994.[8] An accompanying paper in this issue applies some of the methods described to the Cancer Registry of Norway.[9]

### 1.1. *Key issues in the evaluation of data quality*

The three dimensions of quality addressed in the earlier publication[8] were comparability, completeness and validity of registry data. Here, we also consider an additional quality indicator – the timeliness of registry procedures.

*Comparability* of the statistics generated for different population groups (registries, geographical areas, etc.), and over time, is essential to their meaningful interpretation. A basic requirement is the standardisation of practices concerning classification and coding of new cases, and consistency in basic definitions of incidence, such as rules for the recording and reporting of multiple primary cancers occurring in the same individual.

*Completeness* is the extent to which all the incident cancers occurring in the population are included in the registry database. Incidence rates and survival proportions will be close to their true value if maximum completeness in case-finding procedures can be achieved.

*Validity* or *accuracy* refers to the proportion of cases in the registry with a given characteristic that truly have that attribute, and depends on the precision of source documents and the level of expertise in abstracting, coding and recoding.

*Timeliness* of reporting of cancer registry results is an aspect of registry quality that can be considered as a separate issue, although this clearly influences the extent to which data are complete and accurate. Access to recent data is perceived as a priority by users, but, since registries are constantly updating their database as reports are received, and some notifications arrive long after the case was diagnosed, statistics for the recent periods will be incomplete, and will need future updates. There is, therefore, some conflict between the requirement for timely data, and other aspects of data quality, particularly completeness.

In Part I of this review, we consider in more detail the factors influencing comparability, validity and timeliness of cancer registry data, and the present methods available for evaluating the latter two. Part II reviews the topic of completeness.[40]

## 2. Comparability

Comparability is the extent to which coding and classification procedures at a registry, together with the definitions of recording and reporting specific data items, adhere to agreed international guidelines. The comparability of cancer data is thus identified via a thorough review of the registration routines in place, including a specification of the standards and definitions that have been followed. Precise knowledge of current and historical registration practices and definitions are of great importance in the analyses of geographical and temporal variations in cancer incidence. In the evaluation of the comparability of registry data, four topics demand particular attention:

- The system used for classification and coding of neoplasms;
- the definition of incidence, i.e. what is defined as a case, and what is the definition of the incidence date;
- the distinction between a primary cancer (new case) and an extension, recurrence or metastasis of an existing one;
- the recoding of cancers detected in asymptomatic individuals.

### 2.1. *International standards for classification and coding of neoplasms*

The standards of classification and coding of disease are those of the International Classification of Disease, published, since the late 1940s by The World Health Organisation (WHO). A historical overview of the coding system used for cancers is given in Ref. [10] (Fig. 1). For neoplasms, the International Classification of Diseases for Oncology (ICD-O) provides for the coding of

- *Topography* – location of the tumour in the body (synonyms are anatomical location or site);
- *Morphology* – microscopic appearance and cellular origin of the tumour;
- *Behaviour* – whether the tumour is malignant, benign, *in situ* or uncertain;
- *Grade* – the extent of differentiation of tumour;
- A standard coding scheme is also provided for recording the basis of diagnosis of cancers.

### 2.2. *Incidence date*

Cancers develops through an accumulation of mutations in genes. As the natural history of cancer often takes decades from the first mutation to clinical diagnosis, one needs a common definition of 'cancer', from the point of view of deciding whether to register the case, and of the actual date when the disease became incident, from which survival should be measured. Cancers are malignant tumours, and malignancy, although originally a clinical concept, is defined by pathologists in terms of the extent of invasion of the tumour (beyond the basement membrane, for carcinomas). Many studies attest to the fact that there is not complete concordance between pathologists in the diagnosis of cancer.[11] But, finding evidence of 'invasion' will also depend, in part at least, on the diligence with which it is sought in histological specimens. Thus, tumours of the bladder, in which malignant cells are observed at histology, are often considered cancer even though invasion beyond the lamina propria is not observed, since such tumours are often multifocal, with a strong potential for invasion, even if this is not observed in a particular

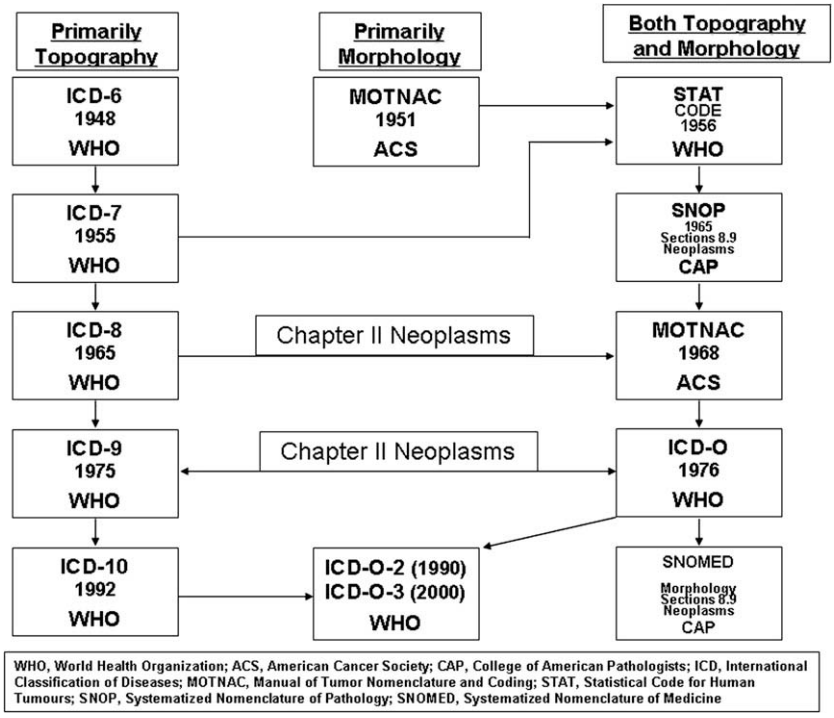


**Fig. 1 – Evolution of the international classification of system for cancers (site and morphology).**

specimen, or section. Statistics on the incidence of bladder cancer are, therefore, particularly prone to non-comparability of definition.[12,13]

As to the date of incidence, various algorithms have been suggested to assist in the choice of the various possibilities available.[14,15] The recommendations of the European Network of Cancer Registries[16] are reproduced in Table 1.

### 2.3. *Multiple primaries*

The fact that it is possible for individuals to develop more than one cancer means that a distinction must be made between those that are new cases (and counted as 'incident' cancers) and those that represent an extension, recurrence or metastasis of an existing one. About 17% of new cancers in Sweden in 2006 were 'second' (or higher order) cancers in the same individual.[17] Cancer registries have adopted various rules for this purpose, but nowadays there are two sets that are more widely used. The rules of the Surveillance Epidemiology and End results (SEER) Programme[15] are used mainly by North American cancer registries[18] whilst internationally, it is the rules developed jointly by the International Association of Cancer Registries (IACR) and the IARC[19,20] that are used, at least for the purposes of reporting incidence rates. The SEER rules result in rather higher incidence rates, since they allow for new (incident) cancers at the same body site, if they occur two months to five years (depending on site) after an earlier diagnosis, whilst the IACR/IARC rules permit only one cancer per body site in a lifetime (unless of different histological types). The SEER rules also allow for the counting of new cases at different subsites of the same organ (e.g. the colon and the skin), or on the opposite side (for paired organs). Fig. 2 shows time trends in the age-standardised incidence rates of breast cancer in the SEER registries 1973–1997, calculated using these different rules.[21] The current IACR/IARC rules are available at http://www.iacr.com.fr/MPrules_ju-

**Table 1 – Standards recommended for the definitions of incidence given by the European network of cancer registries (ENCR, 1999).**

| | Rules for registration of incidence date, in decreasing order of priority: |
|---|---|
| 1 | Date of first histological or cytological confirmation of this malignancy (with the exception of histology or cytology at autopsy). This date should be, in the following order:<br>Date when the specimen was taken (biopsy)<br>Date of receipt by the pathologist<br>Date of the pathology report |
| 2 | Date of admission to the hospital because of this malignancy. |
| 3 | When evaluated at an outpatient clinic only: date of first consultation at the outpatient clinic because of this malignancy |
| 4 | Date of diagnosis, other than 1, 2 or 3 |
| 5 | Date of death, if no information is available other than the fact that the patient has died because of a malignancy |
| 6 | Date of death, if the malignancy is discovered at autopsy |
| Note: | Whichever date is selected, the date of incidence should not be later than the date of the start of the treatment, or decision not to treat, or date of death. The choice of incidence does not determine the coding of the item 'basis of diagnosis' |

ly2004.pdf, whilst the SEER rules can be found at http://seer.cancer.gov/tools/mphrules/.

### 2.4. *Incidental diagnosis*

Incidental diagnosis refers to the detection of cancer incidentally, in asymptomatic individuals. This may result from microscopic examination of tissues that have been removed for a non-cancerous condition, for example, systematic histological examination of material removed at transurethral prostatectomy was responsible for the detection of many 'incidental' (non-symptomatic) cancers of the prostate in the US.[22,23]

Two particularly frequent sources of incidental diagnosis are the detection of cancers as a result of a screening examination, or at autopsy.

#### 2.4.1. *Screen-detected cancers*

Screening aims to detect cancers that are asymptomatic, at an earlier stage than those presenting clinically, hence permitting a greater probability of cure. When a screening programme is introduced into a population, incidence rates increase because the test identifies 'prevalent' cancers that are detectable, but have not yet progressed to the stage where they cause symptoms. The cancers are diagnosed at an earlier age, so that, for cancers for which incidence rates rise with age, the age-specific incidence curve is displaced to the left. After the initial rounds of screening, the 'prevalent' cases will all be detected, and incidence falls, but not to its pre-screening level, because of some degree of 'overdiagnosis' The latter term refers to the detection of cases that would otherwise never have been diagnosed during the subjects lifetime, either because he/she would have died from some other cause in the meantime, or the 'cancer' so found would never have become invasive ('pseudodisease'). Overdiagnosis is a particular problem of screening for prostate cancer. The PSA test identifies many small latent tumours, most of which will not progress to invasive cancer, and it is impossible – at present – to distinguish which of those detected by a PSA test will do so. Incidence rates in the presence of screening can, therefore, be very high.

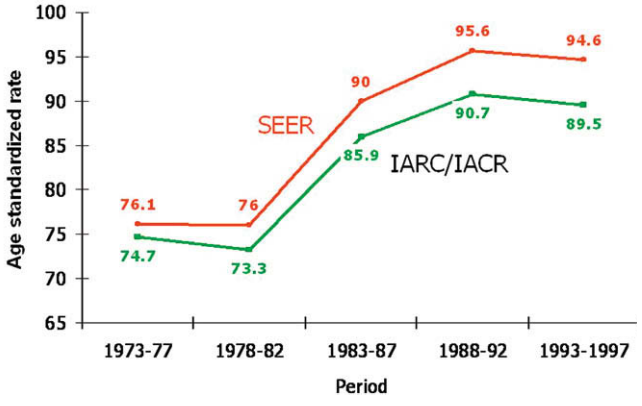


**Fig. 2 – Comparison of trends in breast cancer incidence rates in SEER registries using multiple primary rules of SEER and IARC/IACR, 5-year aggregates 1973–1997.**

It is very useful to be able to distinguish which cancers have been detected by screening in order to assess their impact on trends in incidence rates. Screening, for example, will tend to reduce incidence rates of colon and cervix cancer via the removal of premalignant lesions. The second reason is to enable estimation of the rate of 'interval cancers' – those occurring between screening rounds – as a means of evaluating the effectiveness of the programme.[24] Ideally, this can be done by the linkage of the screening programme and the cancer registry. This allows separation of cancers in non-respondents or non-invited individuals. For this purpose, the ENCR has recommended that cancer registries should collect a data item called 'method of detection in relation to screening'.[25]

#### 2.4.2. *Autopsy diagnosis*

Autopsy diagnosis occurs when cancer is diagnosed through autopsy, without any suspicion of malignancy before death. Evidently, rates are particularly high in countries and regions where legislation enables autopsies to be done for medical, scientific or educational reasons without consent.

Autopsy studies have found high percentages of previously undiagnosed ('latent') cancers of the prostate: over half of men aged over 75 in some areas.[26,27] Thus, where autopsies have been frequently performed, the incidence of cancers that are prevalent in latent form can be dramatically altered. Saxén drew attention to this by comparing the incidence of prostate cancer in Malmo, Sweden (where, in 1972, 46% of cases had been diagnosed at autopsy alone) with the incidence in the country as a whole (where only 7% were diagnosed at autopsy).[12] Rates in Malmö were double those in Sweden.

## 3. Validity (accuracy)

Validity is defined as the proportion of cases in a dataset with a given characteristic (e.g. site and age) which truly have the attribute. Parkin et al.[8] describe four groups of methods that provide numerical indices of validity, at least on an interval scale, and hence permit comparisons with other registries, or, within a registry, over time, or with respect to specified subsets of cases (certain sites, from different sources, etc.). They are: reabstracting and recoding, diagnostic criteria methods (histological verification, and death certificate only), missing information analyses, and internal consistency methods.

### 3.1. *Reabstracting and recoding*

Reabstracting and recoding audits are used to evaluate validity (agreement with source medical records), and reproducibility (agreement amongst data collectors) of registry data. Audits involve samples of cases and are performed in accordance with a study protocol that states the study objectives, describes the sampling scheme and outlines plans for the analysis. These methods provide the most objective way of assessing the validity of cancer registration.[5]

*Reabstracting* describes the process of independently reabstracting records from a given source, coding the data, and comparing the abstracted and coded data with the informa-

tion recorded in the registry database. The objective is to characterise the level of agreement between data in the registry and data reabstracted and recoded from source records (usually hospital medical records) by expert auditors. For each reabstracted data item, the auditor's codes are compared to the original codes to identify discrepancies. If the codes do not match, the discrepancy is classified as to severity. A variety of grading schemes for severity have been used.[8,28] These studies require an arbitration or reconciliation mechanism to determine which of the discrepant answers is correct for the purpose of the study.

*Recoding* involves independently reassigning codes to abstracted text information (without reviewing the source documents), and evaluating the level of agreement with records already in the registry database. As in a reabstracting study, for each recoded case, codes for each data item are compared for discrepancies with those assigned by the expert. This type of study is useful in training new coders; it is easier and less expensive to perform than reabstracting, but the method cannot detect problems with abstracting.

*Reliability* studies are designed to test participants' understanding and adherence to coding rules and practices. This type of study evaluates the overall performance of coders and abstractors. The participants code from identical source documents under controlled conditions. When the coding phase of the study is complete, the coders and abstractors can work with experts to reconcile answers. The final results can be statistically represented by comparing the results to accuracy goals for each data item.

### 3.2. Histological verification

For most cases, the accuracy of the stated diagnosis is likely to be higher if it is based on histological examination by a pathologist. Many cancer registries code diagnoses based on exfoliative cytology or on haematological examination of peripheral blood in the same category as histological examinations, so that it is impossible to distinguish between them. For this reason, the index of validity generally is the percentage of cases *morphologically* verified.

The main value of the percentage morphologically verified (MV%) is an indicator of the validity of the diagnostic information.[8] However, a very high proportion of cases diagnosed by histology or cytology/haematology – higher than might reasonably be expected – is suggestive of an over-reliance on the pathology laboratory as source of information, with failure to find cases diagnosed by other means, and consequent incompleteness of registration.

The absolute value of the MV% can have little meaning, without comparison with an 'expected' value that is reasonable given the circumstances (state of medical technology, local clinical practice) in which the registry operates. Therefore, the MV values (usually calculated by site and sex) should be compared with an appropriate set of standards, so that the values that are significantly different can be identified. Standards may be derived based on the average from registries in the same region (by sex and site) (as in *Cancer Incidence in Five Continents* (CI5), Volume VIII,[29] or in the same country, if the MV values are being examined for sub national areas. A suitable test has been described by Parkin and Plummer (Appendix).[30]

### 3.3. Death certificate only (DCO)

The proportion of cancers for which no other information other than a death certificate mentioning cancer can be obtained – death certificate only registrations or DCOs – is another measure of validity, since the information on death certificates is well known to suffer from lack of accuracy, or lack of precision, compared with that obtained from clinical or pathology records. It is often (mistakenly) cited as an indicator of completeness.[31,32] DCO cases represent the residuum of cases – *after all trace-back manoeuvres have been completed* – for which no other information other than a death certificate mentioning cancer could be obtained (Fig. 3). Thus, whilst a high percentage of DCOs is indicative of inadequate case-finding procedures, a low value cannot guarantee the converse. Since most registries attempt to trace-back cases notified to them via death certificates – the so-called death certificate notified (DCN) cases – a low DCO% may simply reflect success in tracing the cases missed by the normal case-finding procedures. Death certificate initiated (DCI) cases are distinct from DCN cases in that they comprise cancer cases registered after the trace-back procedures have been completed, either from a source identified though the trace-back procedures, or as DCO cases – those for which an additional source cannot be found (Fig. 3).

Establishing acceptable and objective criteria of DCO% have been a contentious issue in international comparative studies. There are many considerations involved in the interpretation of the DCO%, since it is well known to be sensitive to local circumstances, e.g. availability of death certificates, success in record linkage to the registry database, quality of cause-of-death statements, and facility to trace-back cases. Despite this, categories of data quality have been defined in CI5 Volume IX[33] based, in part, on this criterion. However, as for other indices (such as the M:I ratio and MV%), local values can be compared with a standard, based on other registries in the same country, or within a region where similar practices with respect to death registration and the use of death certificates within the registry apply. *Cancer in North America*[34] for example, includes only cancer registries for which the percentage of all cases derived from DCOs is less than 5%.

### 3.4. Missing information

The proportion of registered cases with unknown values for various data items can be an indicator of data quality. Unknown values can result from problems with:

- the data collection system, or access to necessary source documents,
- item and code values that are defined,
- misapplication of coding rules.

Unknown values may, however, have nothing to do with the registration process, but reflect inadequate case histories

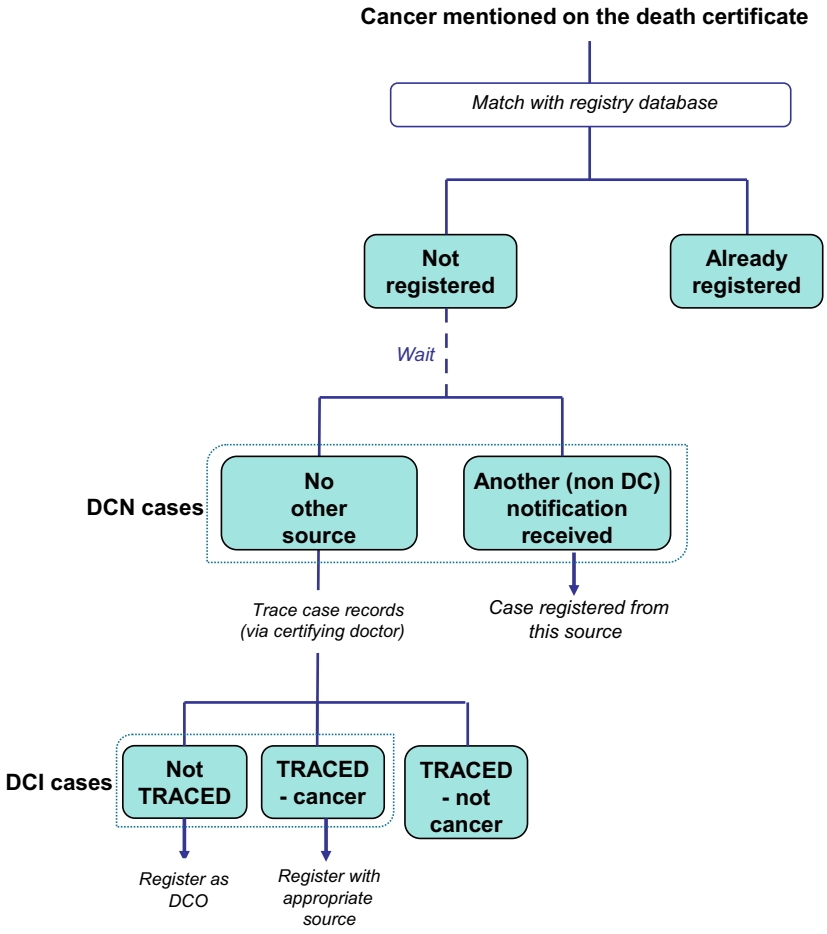


**Fig. 3 – The use of death certificates to identify new cases of cancer.**

or investigation, or ambiguity in the medical record. The definitions used will influence the proportion of unknown codes, for instance, when evaluating cases with *primary site uncertain* (PSU). Other variables for which the proportion of cases with missing values are commonly evaluated include age, ethnicity and stage. The proportion of unknown values usually varies by primary site and tends to be greater amongst elderly patients.

The precise composition of the PSU category varies somewhat between registries. The ICD-O includes a rubric for 'Unknown Primary Site' (C80 in ICD-10), but the PSU category may contain other rubrics (e.g. malignant neoplasms of ill-defined organs of the digestive system (ICD-10 C26), respiratory system (C39), endocrine system (C75), and peritoneal and retroperitoneal neoplasms (C48) as well as those of 'Other and Ill-defined Sites' (C76). A high proportion of cases assigned to the PSU category generally implies poor diagnostic precision (as evidenced by the low MV% observed for this rubric), or failure to specify the site of the primary cancer in cases diagnosed on the basis of tissue obtained from a metastasis. Incidence rates for cancers at specific sites will be underestimated if a significant proportion of registered cases appear in the 'Otherwise and/or Unspecified' category, rather than with their true diagnosis. Cancer in North America includes only cancer registries for which fewer than 3% of cases have missing information for sex, county of residence at diagnosis, and age at diagnosis, and fewer than 5% of the cases had an unknown race.[34] CI5 Volume IX also incorporates acceptable maxima for the percentage of cases with age unknown (<20%), ill-defined sites (<20%) and unknown basis of diagnosis (<20%), although the authors consider the latter to be an indicator of completeness.[33]

### 3.5. *Internal consistency*

Data edits are logical rules, typically embodied in a computer algorithm, that evaluate to 'true,' 'false' or 'maybe', for any value(s) of (a) data item(s). Edits are applied to all records to check for item validity, internal consistency and inter-record consistency. Data edits may involve a single field, multiple fields in a single record, multiple fields in different records within one database or multiple fields in multiple databases. Data editing may be performed interactively, or as a batch process, or both, and can be carried out at several points in the registration process, for example:

- to newly submitted tumour records before they are linked with the central registry database;
- to the registry database, after the addition of new records;
- to the registry database, after any changes have been made.

The check/edit procedure should allow for 'override flags', for those records which the edit identifies a rare condition that needs review, but which have already been verified to be correct. The override flag prevents the condition from continuing to be identified as an error in subsequent runs of the edit/check procedure.

A programme prepared by IARC, *IARC-CHECK*[35] carries out a variety of edits for data coded to ICD-O-3. This includes verification of the validity of codes for sex, incidence and birth dates, and ICD-O-3 topography and morphology codes. It also carries out a series of checks for the validity and for the consistency between variables:

- age/incidence and birth dates
- age/site/histology
- site/histology
- sex/site
- sex/histology
- behaviour/site
- behaviour/histology
- grade/histology
- basis of diagnosis/histology

The programme is available at the IACR website http://www.iacr.com.fr.

## 4. Timeliness

Rapid reporting of information on cancer cases is another priority for cancer registries. Speedy access to cancer information is of clear benefit to health providers and researchers, and early provision of data usually enhances the reputation of the registry. There is, however, a trade-off between timely data and the extent to which it is complete and accurate; the speed at which the registry is able to report depends on numerous factors, and some of which are within the registry's control and others that are not.

There is no formal definition of timeliness in a cancer registration context, although in general terms, it relates to the rapidity at which a registry can collect, process and report sufficiently reliable and complete cancer data. Smith-Gagen et al.[36] have gone further, defining the indicator *time to availability* as the interval between date of diagnosis (or *date of incidence*; however, defined) and the date the case was available in the registry for research. For many registries, the latter will equate to the date at which the database is 'frozen' in order to calculate annual statistics for an official report. This period comprises two intervals: the *time until receipt* (time from the date of diagnosis to the receipt of report), and the *process time* (time from the date of receipt until availability).

The transmission of cases from a reporting facility to the cancer registry influences the timeliness of reporting, and registries often have predefined time intervals (e.g. monthly, twice yearly) at which they expect to receive a certain number of notifications. Once the registry obtains the cancer records, there still remains a series of time-consuming process steps to complete, including the retrieval and matching of death certificate notifications to the registry database, and the initiation of trace-back procedures for unmatched cases.

Efficient procedures for the retrieval of cancer reports and the dispatch of reminders and a dedicated and well-trained staff can influence the timeliness of reporting of cancer cases. Data may be transmitted to cancer registries from the same notifying sources (hospitals and laboratories) in 'real-time'. Electronic data capture has expedited case identification and the abstracting process for some registries.

Whilst there are no international guidelines for timeliness of cancer registry data, several North American agencies – notably those that provide funding via contracts – have set out specific standards for the relevant registries:

- *American College of Surgeons*: Cases must be abstracted within 6 months of date of first contact.[37]
- *Surveillance, Epidemiology, and End Results programme*: The registry is under contract to report complete counts of new cases to the USA NCI Within 22 month of the end of the diagnosis year (see http://seer.cancer.gov/faq/).
- *Centers for Disease Control and Prevention/National Programme of Cancer Registries*: Within 12 months of the close of the diagnosis year, 90% of expected unduplicated cases are available to be counted as incident cases at the central cancer registry; and, within 24 months of the close of the diagnosis year, 95% of expected unduplicated cases are available to be counted as incident cases at the central cancer registry.[38]
- *North American Association of Central Cancer Registries*: Within 23 months of the close of a diagnosis year, the registry should contain at least 95% of the expected cases of reportable cancer occurring in residents during that year.[28]

Registries are constantly updating their database as reports are received, and some notifications, especially those from death certificates, arrive long after the case was diagnosed. Incidence figures for the recent years will, therefore, need future updates as incoming data accrue. A delay of some years will usually be of lesser importance in the study of cancer patterns either in terms of incidence at different sites, or comparisons between different places. It is also true, however, that the receipt of data considered 'out of date' generally causes distress outside the registry community.

The issue of timeliness will commonly have a greater impact on the study of time trends of incidence, particularly if the level of undercounting becomes greater with each year up to the latest reported. Several approaches have been proposed to produce more timely statistics, whilst minimising the effects of under-reporting. In the USA, for example, researchers linked with the SEER programme have constructed a delay model[39] that estimates the undercount at the time of reporting, and adjusts the reported rates accordingly. Both unadjusted and delay-adjusted rates, according to a delay of 22 months after the close of the year of registration are available from the SEER Cancer Statistics Review (http://srab.cancer.gov/delay/canques.html).

## Conflict of interest statement

None declared.

## Acknowledgements

This study was undertaken as part of the Cancer Control using Population-based Registries and Biobanks (CCPRB) project, funded by the EU Sixth Framework Programme (FP6-2002-LifeSciHealth Contract No. 503465).

## Appendix. Statistical test for comparison of the registry MV%, by site, with standard values[30]

The comparison populations are registries in the same country, or region, indexed by $i = 1,\ldots, n$.

$y_i$ is the number of microscopically verified cases in registry i and $d_i$ is the total number of cases in registry i.

Under the binomial model for $y_i$

$$E(y_i) = pd_i$$

$$\mathrm{Var}(y_i) = p(1-p)d_i$$

where $p$ is the proportion of MV cases.

The overdispersion model changes this to

Var (yI) = $\phi\varphi(1-p)d_i$

The parameters are estimated by

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} d_i}$$

$$\hat{\phi} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(y_i - pd_i)^2}{d_i p(1-p)}$$

For the registry under test, with data $y_j$, $d_j$, the test statistic is

$$Z^2 = \frac{(y_j - \hat{p}d_j)^2}{\hat{\phi}\hat{p}(1-\hat{p})d_j} \sim \chi_1^2$$

So the registry is flagged as unusual if $Z^2 \geqslant 3.84$.

## REFERENCES

1. Parkin DM. The evolution of the population-based cancer registry. *Nat Rev Cancer* 2006;**6**:603–12.
2. Armstrong BK. The role of the cancer registry in cancer control. *Cancer Causes Control* 1992;**3**:569–79.
3. Parkin DM. The role of cancer registries in cancer control. *Int J Clin Oncol* 2008;**13**:102–11.
4. Storm HH. Cancer registries in epidemiologic research. *Cancer Causes Control* 1996;**7**:299–301.
5. Brewster D, Crichton J, Muir C. How accurate are Scottish cancer registration data? *Br J Cancer* 1994;**70**:954–9.
6. Storm HH, Michelsen EV, Clemmensen IH, Pihl J. The Danish cancer registry – history, content, quality and use. *Dan Med Bull* 1997;**44**:535–9.
7. Skeet RG. Quality and quality control. In: Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG, editors. *Cancer registration. Principles and methods (IARC scientific publications no. 95)*. Lyon: IARC Scientific Publications; 1991. p. 101–7.
8. Parkin DM, Chen VW, Ferlay J, Galceran J, Storm HH, Whelan SL, editors. *Comparability and quality control in cancer registration. (IARC technical report no. 19)*. Lyon: IARC (WHO) and IACR; 1994.
9. Larsen IK, Småstuen M, Johannesen TB, et al. Data quality at the cancer registry of Norway: an overview of comparability, completeness, validity and timeliness. *Eur J Cancer* 2009.
10. Fritz A, Percy C, Andrew J, et al., editors. *International classification of diseases for oncology: ICD-O*. 3rd ed. Geneva: World Health Organization; 2000.
11. Wells WA, Carney PA, Eliassen MS, Tosteson AN, Greenberg ER. Statewide study of diagnostic agreement in breast pathology. *J Natl Cancer Inst* 1998;**90**:142–5.
12. Saxen E. Trends: facts or fallacy. In: Magnus K, editor. *Trends in cancer incidence: causes and practical implications*. Washington, DC: Hemisphere Publication; 1982. p. 446.
13. Crow P, Ritchie AW. National and international variation in the registration of bladder cancer. *BJU Int* 2003;**92**:563–6.
14. Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG, editors. *Cancer registration. Principles and methods (IARC scientific publications no. 95)*. Lyon: IARC Scientific Publications; 1991.
15. SEER. *Multiple primary and histology coding rules (information for cancer registrars: data collection and coding manuals)*. <http://seer.cancer.gov/tools/mphrules/>; 2007.
16. European network of cancer registries. *ENCR RECOMMENDATIONS-recommendations for coding incidence date*. <http://www.encr.com.fr/ENCR.htm>; 1999.
17. Socialstyrelsen. Sweden: the national board of health and welfare, centre of epidemiology. *Cancer incidence in Sweden 2006*. <http://www.socialstyrelsen.se/Statistik/statistik_amne/Cancer>; 2007.
18. Havener LA, Thornton ML. *Standards for cancer registries volume II: data standards and data dictionary*. 13th ed. Springfield (IL). <http://www.naaccr.org/filesystem/pdf/Standards%20Volume%20II,%20version%2011.3%20posted%20copy.pdf>; 2008.
19. Working Group Report. International rules for multiple primary cancers (ICD-0 third edition). *Eur J Cancer Prev* 2005;**14**:307-8.
20. International Association of Cancer Registries. International rules for multiple primary cancers. *Asian Pac J Cancer Prev* 2005;**6**:104-6.
21. Parkin DM. *Standards to ensure quality of cancer registry data. JACR monograph no. 11*. <http://www.cancerinfo.jp/jarc/Pub_m/11_parkin.pdf>; 2007.
22. Potosky AL, Kessler L, Gridley G, Brown CC, Horm JW. Rise in prostatic cancer incidence associated with increased use of transurethral resection. *J Natl Cancer Inst* 1990;**82**:1624–8.
23. Shimizu H, Ross RK, Bernstein L, et al. Cancers of the prostate and breast among Japanese and White immigrants in Los Angeles County. *Br J Cancer* 1991;**63**:963–6.
24. Day NE, Williams DR, Khaw KT. Breast cancer screening programmes: the development of a monitoring and evaluation system. *Br J Cancer* 1989;**59**:954–8.
25. European network of cancer registries. *ENCR RECOMMENDATIONS-method of detection in relation to screening*. <http://www.encr.com.fr/ENCR.htm>; 2007.
26. Breslow N, Chan CW, Dhom G, et al. Latent carcinoma of prostate at autopsy in seven areas. The international agency for research on cancer, Lyons. *France Int J Cancer* 1977;**20**:680–8.
27. Yatani R, Chigusa I, Akazaki K, et al. Geographic pathology of latent prostatic carcinoma. *Int J Cancer* 1982;**29**:611–6.
28. Havener L. *Standards for cancer registries volume III: standards for completeness, quality, analysis and management of data*. Springfield (IL): North American Association of Central Cancer Registries; 2004. <http://www.naaccr.org/>.
29. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. *Cancer incidence in five continents. IARC scientific publications no. 155*, vol. VIII. Lyon: IARC; 2002.
30. Parkin DM, Plummer M. Comparability and quality of data. In: Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB, editors.

*Cancer incidence in five continents (IARC scientific publications no. 155)*, vol. VIII. Lyon: IARC. p. 57–73.
31. Brenner H. Limitations of the death certificate only index as a measure of incompleteness of cancer registration. *Br J Cancer* 1995;**72**:506–10.
32. Crocetti E, Miccinesi G, Paci E, Zappa M. An application of the two-source capture-recapture method to estimate the completeness of the Tuscany cancer registry, Italy. *Eur J Cancer Prev* 2001;**10**:417–23.
33. Curado MP, Edwards B, Shin HR, et al. *Cancer incidence in five continents. IARC scientific publication no. 160*, vol. IX. Lyon: IARC; 2007.
34. Wu XC, McLaughlin CC, Lake A, et al. *Cancer in North America, 2000–2004. Volume one: incidence*. Springfield, IL: North American Association of Central Cancer Registries, Inc. (NAACCR); 2007.
35. 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: IARC; 2005.
36. Smith-Gagen J, Cress RD, Drake CM, Felter MC, Beaumont JJ. Factors associated with time to availability for cases reported to population-based cancer registries. *Cancer Causes Control* 2005;**16**:449–54.
37. American College of Surgeons. *Commission on cancer program standards 2004 (revised edition)*. Chicago: ACS; 2006. Available from: <http://www.facs.org/cancer/coc/cocprogramstandards.pdf>.
38. Centers for disease control and prevention. National program of cancer registries. *FY 1997 cooperative agreement guidance non-competing continuation applications*. <http://www.cdc.go/cancer/nocr/pncrpdfs/426guide.pdf>; 1997.
39. Midthune D, Fay M, Clegg L, Feuer E. Modeling reporting delay and reporting corrections in cancer registry data. *J Am Stat Assoc* 2005;**100**:61–70.
40. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *Eur J Cancer* 2009;45:756–64.