# Appendix 4

# Automation in cancer registration

**Background**

Algorithms aimed at replacing the manual decision-making process, usually carried out by registry personnel on *ad hoc* registry forms, were first introduced in the early 1970s by the Ontario Cancer Registry (OCR). The forms containing the information on cancer patients were coded and computerized at the OCR and subsequently treated by software developed by the OCR for this purpose. All the diagnoses of cancer were assigned by the program using the data available electronically (Clarke *et al*., 1991).

In the early 1990s, the project of the Venetian Tumour Registry (RTV) explored for the first time the possibility of using, as primary sources of data for the registration process, electronic routine data from the hospitals, pathology departments, and other health institutions in the north-east region of Italy.

The data were coded according to ICD-9 (hospital discharges and death certificates) or SNOMED (pathology records). By applying a rather simple algorithm (Table 1) to the electronic data, it was possible to assign a diagnosis to the majority of the incident cases (Simonato *et al*., 1996).

The methodology was subsequently adopted by the Northern Ireland Cancer Registry (NICR), and partially by the Thames Cancer Registry, with similar results. More recently, a network of registries in the north-east of Italy, the North-East of Italy Cancer Surveillance Network (NEICSN) adopted and further developed the registration system.

Basically the method consists of a binary decisional system of concordance/ discordance, through which a potential incident case is accepted with a consolidated diagnosis of cancer, or rejected. Cases rejected by the program are resolved manually by the registry personnel. Figure 1 illustrates the standard data flow, while an example of its application by NEICSN is shown in Figure 2.

The example reported in Figure 2 shows how incidence was obtained by NEICSN for the period 1999–2000. All the electronic data available from the three sources, consisting of 4,401,914 hospital discharges, 197,859 death certificates and 2,516,832 pathology records are used in the process, in which the first phase consists of record linkage with the various sources, in order to eliminate cancer cases with diagnosis before 1 January 1999.

Of these, 1,103,147 (15.4%) records with a diagnosis of neoplasia are selected and *summarized* into 305,369 subjects with at least one record of cancer (average number of records per subject 3.6). Out of these, 246,655 (80.8%) were prevalent cases, and 8,869 (2.9%) turned out to be non-residents at the moment of diagnosis. This leaves 49,845 subjects potentially affected by at least one incident cancer.

In the following step, the cases are *consolidated* by the program, and 39,148 cases are entered in the registry database. These constitute 78.5% of all cancer cases. The remaining 10,697 cases are revised and 21.5% of the total are entered by the registry personnel.

A large and increasing number of cases accepted belong to the categories of benign tumours, *in situ* tumours, and tumours of uncertain nature. This would allow follow-up studies of non-malignant tumours.

**Table 1. NEICSN criteria for case consolidation**

| No. | *The criteria according to which the SITE program operates* |
|---|---|
| **1.** | Cancer cases with full concordance between two or more sources |
| **2.** | Histologically confirmed cases with at least one concordant or compatible (e.g. metastases or ill-defined) hospital discharge or death certificate |
| **3.** | Histologically confirmed skin cancer (ICD 173) unless in combination with skin melanoma (ICD 172) |
| **4.** | Histologically confirmed benign, *in situ*, and uncertain behaviour tumours |

**Applicability of automated cancer registration (ACR) techniques**

The ACR methodology can be used only if the three traditional sources of information for a registry (hospital discharges, death certificates, pathology records) are available in an electronic form and are coded according to the ICD classification. Pathology records, often coded according to SNOMED, are transformed into ICD through a conversion table.

If a registry is just starting operation, it is recommended to wait for a number of years before starting to calculate incidence, in order to avoid the inclusion of prevalent cases in the incidence figure. This problem does not apply to existing registries, which will identify prevalent cases by record-linkage with their historical database.

The completeness and the quality of the original electronic data are crucial in determining the efficiency of the automated process. This needs to be carefully checked before embarking on ACR processing of the data. Low quality of the information sources will increase the proportion of discordant diagnoses, resulting in a lower efficiency of the system, while it is less likely to produce false positives as independent sources have a very low probability of making concordant ICD errors.

**Summarization**

This is the process by which the electronic records containing health information are linked to the individuals in the population file by using an ID code. Once the quality of the original data is ascertained, electronic and coded records undergo a process of record linkage with the population file resulting in a number of individuals for whom one or more cancer diagnoses have been *summarized* by computer.

These cancer histories are then ordered chronologically, which allows the computer-based exclusion of prevalent cases. The remaining individuals are the patients potentially affected by an incident neoplastic disease, who have one or more records with a coded diagnosis of cancer.

**Consolidation**

This is the process by which software based on the algorithm adopted evaluates the consistency of the coded diagnosis of cancer within the same subject, according to a number of established criteria.

There is no standard at present, and one of the goals of the ENCR Working Group on Automated Cancer Registration is to agree on the number and nature of these criteria. Those currently used in the algorithm by the NEICSN are presented in Table 1. Further development is in progress, but the results do not differ greatly between the few existing automated registries.

The proportion of accepted cases ranges from 50% to 75%, the variability being attributable more to the characteristics of the information sources than to the performance of the algorithm.

Certain groups of cases are systematically rejected by the present algorithm and are therefore manually checked by the registry personnel. These represent the largest proportion of rejected cases and comprise multiple tumours (non-melanotic skin cancer excluded), cases based on hospital records only, and cases based on death certificate only.

The system also systematically registers non-malignant tumours (benign, *in situ*, uncertain) which could be of interest for prospective studies of individuals at higher risk.

**Developments**

Management of large databases requires a sophisticated and efficient record linkage system, which implies a variable degree of computer-assisted decision making.

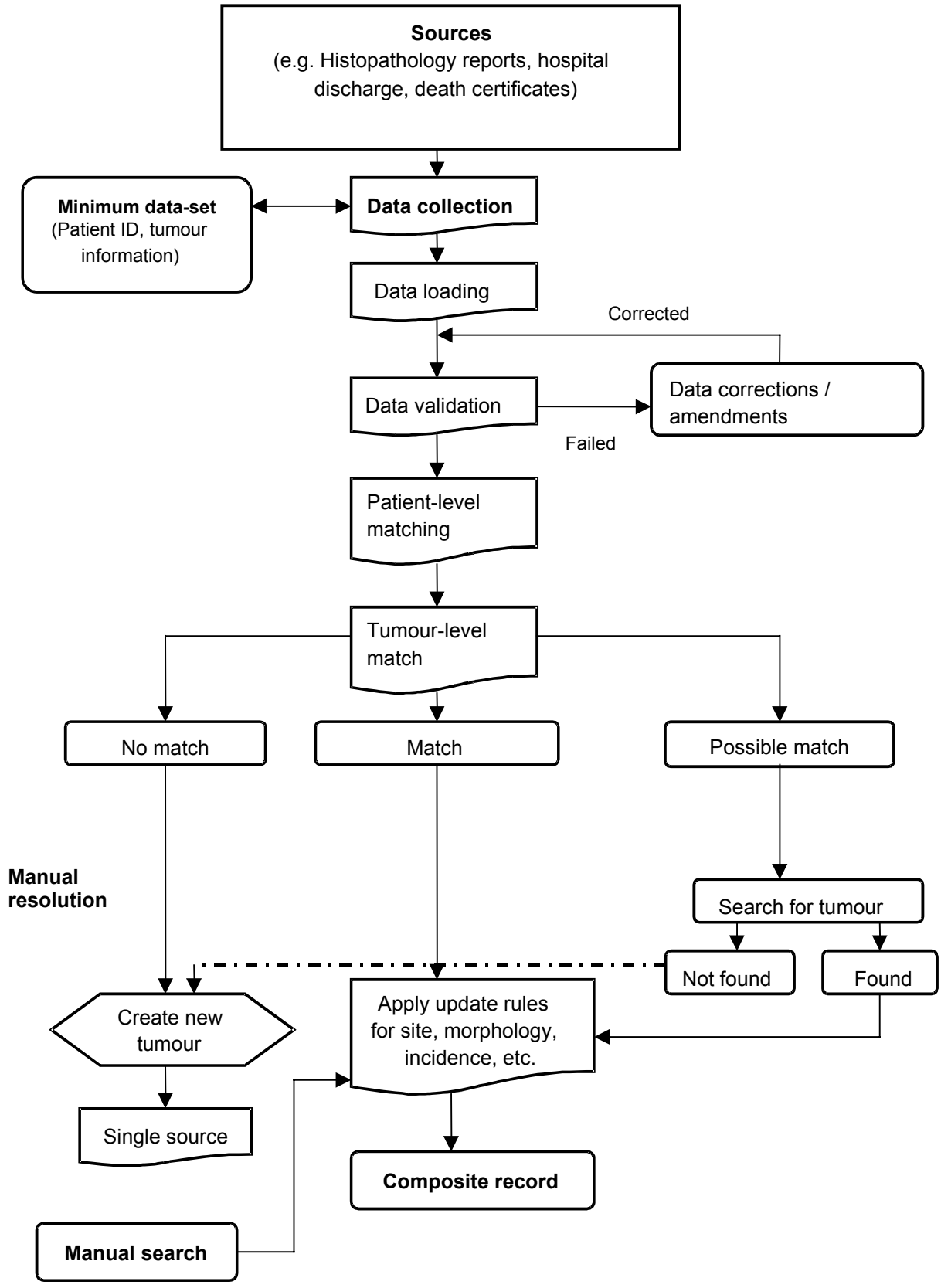
The increasing availability of coded pathology, hospital and death certificate records offers the possibility, not previously available, of directly using coded information for case resolution.

This development is promising, but may introduce additional problems regarding certain aspects of cancer registration which have already been under scrutiny, such as quality of diagnosis, and, even more important, comparability of cancer incidence data across cancer registries. The results so far available do not indicate major new problems of quality from registries which use the ACR methodology. The evidence is, however, at present based on the experience of very few registries and needs further evaluation. Very important is the issue of comparability, both between registries and within the same registration system, when moving from a manual system to ACR.

The availability of computerized data of different quality might lead to considerable differences between cancer registries, particu-

larly when different algorithms for case consolidation are used. This will be a crucial issue in the development of ACR, and highlights the need for standardization of the data source definitions, and of the computer-assisted case consolidation processes, an additional task for international organizations such as ENCR and IACR.

**Figure 1. Process of automated cancer registration**



The field of automated processes applied to medical data is evolving, mainly due to the increasing computerization of information in the hospital. Of particular importance is the extension to laboratory and imaging departments, and the increasing availability of electronic data on drug consumption.

This implies that in the very near future, an increasing amounts of different types of computerized information will be available for entry into the automated process, with the target of building up population-based surveillance systems which can be extended also to diseases other than cancer.

In view of such developments in registration of cancer, as well as of other diseases, cancer registries need to plan extension of their activities beyond the production of cancer incidence statistics: to establishing tools, within public health systems, for cancer surveillance, planning intervention studies and their evaluation, and carrying out etiological investigations taking advantage of the easy access to population-based information.

ACR techniques can make a valuable contribution to the development of this new situation by improving the timeliness and completeness, and by reducing the costs, provided that, in parallel, strict and efficient quality control is systematically performed.

## References

Black, R.J., Simonato, L., Storm, H.H. & Démaret, E. (1998) *Automated Data Collection in Cancer Registration* (IARC Technical Report No. 32), Lyon, IARCPress

Clarke, E., Marrett, L.D. & Kreiger, N. (1991) Cancer registration in Ontario: a computer approach. In: Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S. & Skeet, R.G., eds, *Cancer Registration. Principles and Methods* (IARC Scientific Publications No 95), IARC, Lyon

Simonato, L., Zambon, P., Giordano, R., Guzzinati, S., Stocco, C., Tognazzo, S. & Winkelmann, R. (1996) A computerised cancer registration network in the Veneto region, North-east of Italy: a pilot study. *Br. J. Cancer*, **73**, 1436–1439

**Figure 2. Generation of incidence data by the North-East of Italy Cancer Surveillance Network (NEICSN) 1999–2000**