Chapter 8. Manual and computerized cancer registries

R.G. Skeet

Herefordshire Health Authority, Victoria House, Hereford HR4 0AN, UK

Introduction

Superficially, there may appear to be little in common between a small, manual cancer registry dealing with perhaps a few hundred new cases a year with their details in a box file, and a large, highly computerized registry which apparently consists of visual display units (VDUs) and little else. To describe the operation of registries at these two extremes in one chapter may seem inappropriate. A detailed study of both, however, reveals that their functional components are identical. The same basic tasks have to be performed in each—it is only the methods which differ. The nature of modern computing systems is such that it is not always easy for the newcomer to appreciate what actually is being achieved. When cases have been identified (see Chapter 5), the activities in the cancer registry are universal—they are primarily concerned with getting data ready for tabulation and analysis. In this chapter, the operation of both manual and computerized registries will be outlined function by function in order to describe both the manual tasks themselves and the various computer solutions available. The concern is more with concepts and principles that apply to cancer registration than with a description of procedures in one or more prototype registries. Descriptions of the operations of four different registries are given in Appendix 3.

Operational tasks of the cancer registry

In some way or other, every registry must carry out the tasks outlined below. The amount of resources channelled into each will depend upon many factors, often external to the registry itself.

Data collection

No registry can operate without some mechanism for data-gathering. This has been considered in detail in Chapter 5.

Record linkage

Frequently the registry will receive records relating to an individual patient from more than one source—for example a hospital, a pathology laboratory and an office of vital statistics. These records must all be linked to the same patient so that the details of each patient are complete and there are no duplicate registrations for the same tumour. The linkage is a crucial operation, the importance of which cannot be overemphasized.

Data organization

Data for scientific study must be held in an orderly manner. Information arrives at the registry in a more or less structured format—partly on well-designed forms created specifically for the purpose and partly on other reports of a more descriptive nature and designed primarily for other purposes. Computerized data will come to the registry in an already processed, or partly processed, form but it is likely that further organization of the data will still be required in the registry.

Medium conversion

Even in a manual registry, it is unlikely that the information will be retained entirely on the original documents. In the computerized registry, information on paper will have to be transferred onto a machine-readable medium, punched cards, magnetic tape or disk. The computerized registry may hold its data on more than one medium.

Enquiry generation and follow-up

Frequently, the acquisition of an item of information alerts the registry to the fact that information it already has may be incomplete or incorrect. For example, the arrival of a death certificate carrying a diagnosis of malignant disease relating to a recently deceased patient who is not already registered indicates the possibility that the registry has failed to acquire information at an earlier stage. The registry must then make further enquiries in an attempt to obtain full details or to resolve any inconsistencies. Many registries regard the follow-up of their patients as one of their most important functions, and this may take an active or passive form. Active follow-up involves routine periodic requests for further data about registered patients.

Data analysis

The analysis of cancer registration data is considered in Chapters 10–12 but is mentioned here for the sake of completeness and to emphasize that, without this final operation, the preceding tasks are pointless.

The processes described above will be discussed in turn below. While details will be given where appropriate, because cancer registries differ a great deal in their methods of operation, attention will be directed to the main principles involved. No attempt will be made to describe how the tasks should be done in absolute terms, since there is no single solution to similar problems encountered by different registries.

Record linkage

Multiple reports

Before considering the problem of record linkage, it is important to understand the basic concepts of multiple notification, multiple tumour, and duplicate registration.

Multiple notifications

These refer to reports received about a single tumour in one cancer patient. If a patient is diagnosed as having cancer in one hospital and referred to another for treatment, it may well be that both hospitals report the case to the registry. The registry must recognize these reports as multiple notifications.

Multiple tumours

Sometimes a cancer patient develops more than one primary tumour and it is customary to make an independent registration for each, since cancer registries actually count the numbers of primary cancers rather than the number of cancer patients. It is important for a registry to have a clear definition of what constitutes multiple malignancy, to avoid both over- and under-registration of primaries. The study of multiple malignancy is important in its own right (see Chapter 3). A definition of multiple tumours, suitable for international use, is given in Chapter 7 (p. 78).

Duplicate registration

This occurs as a result of a failure in the linkage process, such that a tumour is counted more than once by the registry.

The linkage process

The purpose of record linkage in cancer registration is to bring together records that pertain to the same individual in order to determine whether a report concerns a tumour (case) that is already known to the registry or a new primary tumour.

In its simplest form, this is illustrated in Figure 1. First, the patient is identified as being either unknown to the registry, in which case a new registration is made, or known, in which case a new registration is made only if the notification refers to a different primary tumour. The same basic process applies to both manual and computerized linkage systems. When a new registration is made, of either a new patient or a new tumour, a number is issued by the registry. This is usually referred to as the accession or registration number.

Manual linkage

The purpose of compiling a register as such, i.e., a list of names, is that each new patient can be checked against the list to ascertain whether he or she is already known to the registry. Apart from those countries where a personal identity number is used, the examination of each new name against the register is the only method available for performing this task.

The name alone is usually insufficient, since its discriminatory power may be limited. In the case of a very common name, it is extremely low. The date of birth is included in the linkage process by most registries, since this increases the discriminatory power a great deal. If the name and date of birth were recorded with unfailing accuracy on every occasion, it is unlikely that any other items would be necessary for accurate record linkage. However, this is not so, and most registries will

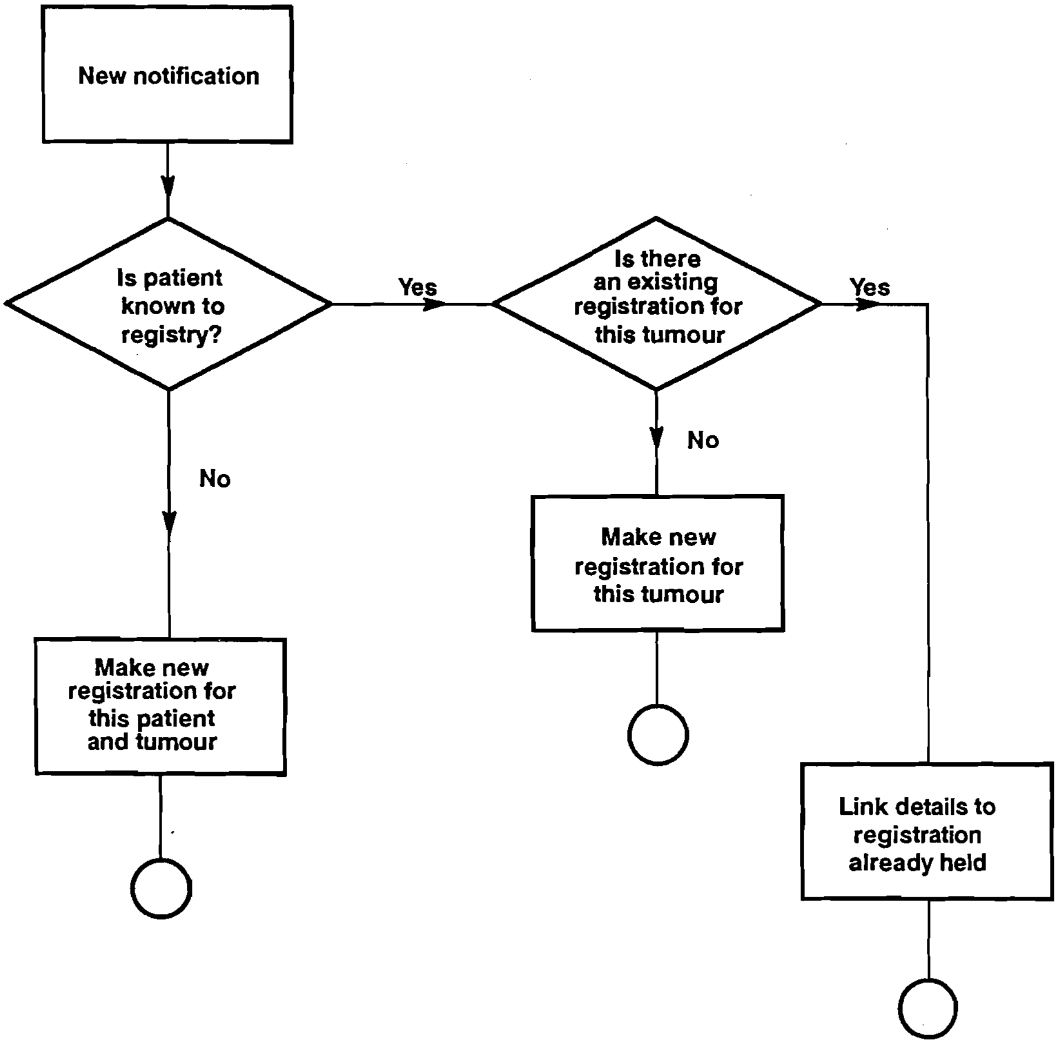

New notification
Is patient known to registry?
Yes
No
Is there an existing registration for this tumour
Yes
No
Make new registration for this patient and tumour
Make new registration for this tumour
Link details to registration already held

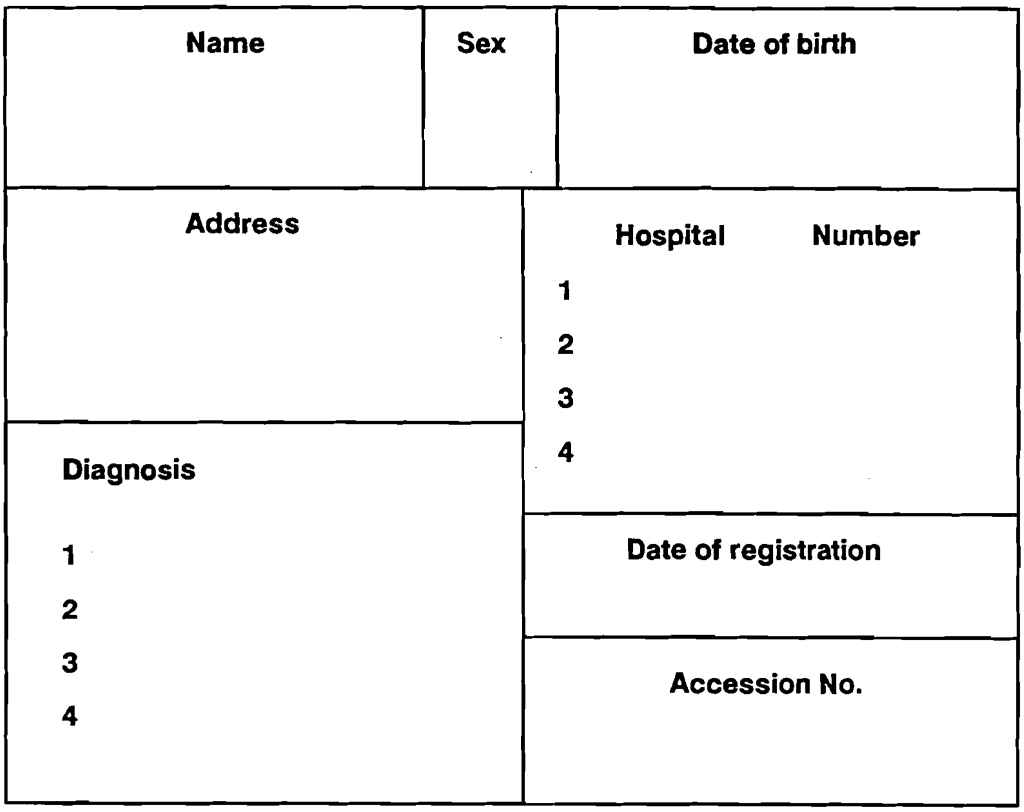
Figure 1. The basic process of linkage in cancer registries

use the patient's address and possibly maiden name also to improve the quality of the linkage.

The traditional method of record linkage is to maintain a file of patient index cards similar to the example shown in Figure 2. All new documents coming to the registry are checked against this index and, as a result, are divided into two groups, depending upon whether a match is found. Where a match is not found, a new accession number (see below) is given to the case and a new patient index card is prepared and filed in the index.

Generally this process is carried out as a batch procedure, usually on a daily, weekly or monthly cycle. The incoming forms are sorted alphabetically according to patient name and then the index is searched, maybe dividing the work between several clerks, each using a different part of the index. Alternatively, the forms may be sorted by birth date.

After the new cases have been identified and numbered, the patient index card is typed and filed. This filing process actually corresponds to a second search of the index and it is, therefore, good practice for this to be carried out by a different clerk. For example, if one clerk searches the first half of the alphabet and another the



Name

Sex

Date of birth

Address

Hospital Number

1

2

3

4

Diagnosis

1

2

3

4

Date of registration

Accession No.

Figure 2. A typical patient index card

second, their roles should be reversed for the filing process. Inadvertently missed matches may be detected in this way and errors corrected.

It is absolutely essential that the filing of the patient index cards is completed before the checking of the next batch begins. Multiple notifications relating to the same patient often arrive at the registry within a short space of time, and much labour can be wasted in searching in the index for a patient whose notification has already arrived and whose index card is awaiting filing.

A major difficulty faced by clerks is that names may be spelled inconsistently on various documents. The author's own name, Skeet, may qualify for some sort of record in this respect—Skeat, Skete, Skate, Sheet, Street, with their plural forms also, are frequently used as a result of mishearing or miscopying.

The nearer the front of the name the error occurs, the greater is the chance of a duplicate registration being set up, and this is not a problem confined to users of the Roman alphabet. The only satisfactory solution is to file the cards in some sort of compromise between a purely alphabetical system and a phonetic system, where names which sound alike are filed together. Thus Symonds, Simmons, Simons, Symon and Simon would all be filed together in the index, perhaps under 'Simmons'. Guide cards are inserted at the appropriate places for other spellings to ensure that the clerk searches correctly. The searching of manual indexes can be developed to a considerable art in which experience plays an essential role. It is well known that the experienced clerk will be able to find names in the index which the Director of the registry will not!

The use of names is a product of the local culture. The adoption of the husband's family name by a woman at marriage is common in many cultures but, even in those, exceptions occur and appear to be on the increase. Names may change for other reasons and abbreviations or nicknames may also be used. Where this is a serious problem, cross-indexing may be helpful but, in any event, a registry should construct its patient index file in accordance with local custom.

Names must never be removed from the alphabetical index. If a patient changes name, both names should be maintained in the index. This means that a second card is created for the same person (sometimes referred to as an also-known-as or 'aka' card), which refers the searcher to the original card. For patients with multiple malignancy, either one index card containing all their diagnoses is kept, or a single card is used for each primary cancer, perhaps stapled together for convenience.

As the patient index grows, the proportion of cards corresponding to dead patients increases. After a period of, say, three years from the date of death, it is unlikely that any further new reports will be received and consideration should be given to transferring the index cards for these patients from the main patient index to a subsidiary dead file, which will be used less frequently. Although this represents much work, in a large registry, prime office space may be at a high premium and, on the basis of storage space alone, this separation may become essential. Removal of what is essentially inactive material from the main index results in a much smaller file, which is easier to use and in which fewer errors will be made.

Computerized linkage

Two main types of computerized record linkage can be found in cancer registries, broadly falling into offline and online categories.

Offline record linkage consists of submitting a batch of prepared records to the computer on disk or magnetic tape. The computer then compares the identifying information on the new records against records already in the system. Various techniques are used to establish the degree of matching of each new record. This may include some method of scoring such that an exact match of name achieves a higher score than a near match, while the absence of any match with an existing name results in a zero score. Similar scores are computed for matching on date of birth. Other data items may be used for comparison, scores being calculated for each, and a final weighted score is then computed. The score is then evaluated—above one critical level the match is assumed to be correct, below another, the absence of a match is assumed. Between these scores fall those pairs of records where a match is a possibility. These are usually printed out in full for manual scrutiny so that a clerk can make a decision on each. These techniques are particularly useful where no further clerical effort is required in processing the data, for example, in dealing with computer files from hospitals which feed a central registry. Matched records automatically update the existing records and unmatched records set up new registrations. This type of registration scheme is used by the Ontario Tumour Registry and is described in detail in Appendix 3(*c*).

Online record linkage is most useful when paper documents are being entered into a computer system using a VDU. Before entering the data themselves, the operator types in the name, date of birth and any other details required for the linkage. The computer then searches its files for cases with the same or similar details and displays possible matches on the screen. On the basis of these, the operator decides whether the case is actually already known to the system or represents what will become a new registration. These systems can have elaborate methods for identifying possible matches using various phonetic procedures. They may also offer considerable flexibility, since the interrogation may take different forms. For example, the date of birth may be fixed and the computer asked to display the names of all cases beginning with a given sequence of letters; alternatively, the name may be given and all cases displayed irrespective of the date of birth. Such record linkage can be extremely fast, since the speed of access to a record using a carefully designed index is not directly related to the number of cases known to the system. With a well structured index, which may be quite complex, it is possible to find an exact match in a file of over a million records in under one second. This depends upon having sophisticated computer programs using data-base management techniques and, in the case of a large registry, a great deal of disk storage which is permanently online. Although looking very different from the manual method, the principle is exactly the same. Instead of the index being stored on cards, it is held on disk, while the software takes the place of the human searcher who knows in which drawer a card will be found if it is present at all. The index file is, of course, maintained by the computer itself. As soon as a case is identified as new the computer automatically sets up an index record, thus eliminating the need to work in batches. Amendments to names or dates of birth can automatically be fed back into the index without deleting the original entries. It is also unnecessary to sort the incoming documents into alphabetical order and the entries for dead cases need not be transferred to another file.

Accession numbering

In most systems, particularly computerized ones, it is convenient to store the data numerically rather than alphabetically. New patients are given a patient registration number or accession number (see Chapter 6) as soon as the linkage process has identified them as such: In the most widely used numbering system, the first two digits signify the anniversary year (however this is defined; see below) and these are followed by a number allocated serially as cases are registered with that anniversary year. Hence, the first case registered with its anniversary date in 1987 would be numbered 8700001, the second 8700002 and so on. The year of registration may be different from the incidence year. During 1987, cases diagnosed in 1985 and 1986 will, no doubt, be registered. These will have the 87 prefix allocated, although the year for calculating incidence will be 1985 and 1986 respectively.

A complication arises in the numbering of multiple malignancies in the same individual. There is much to be said for having one accession number per patient and adding a suffix for tumour number. This makes the linkage between multiple primaries easier and facilitates follow-up. The alternative is to issue more than one accession number to patients with multiple tumours and to supply cross-indexing data in each primary's record but this procedure is not recommended.

An alternative, which is useful for registries with online computer systems, is to issue a patient number to each new cancer patient using the sequence of accession numbers. Each tumour is given a tumour number, the first being the same as the patient number. If another primary is registered in that individual, the same patient number is used but a new tumour number is allocated (the next in the accession sequence). Data are stored and processed using the tumour number, but a patient's various primary cancers can be linked together because they have the same patient number.

Confidentiality

Cancer registration today is carried out against a background of growing concern over the confidentiality of personal data. For all registries it is absolutely essential that enough details are obtained to identify each patient for, without them, it is impossible to link multiple notifications including those coming by way of a death certificate. For the vast majority of registries this means having the name, and probably the address, of each case. Without the ability to distinguish one patient from another, the cancer registry cannot operate. The matter of confidentiality is considered further in Chapter 15.

Data organization

A separate record is created for each registered primary tumour; thus, a patient with multiple primary tumours will have multiple tumour records. It is recommended that a special code is used to indicate the presence of multiple tumours (see Chapter 6). The items which could be contained and coded in the tumour record are described in detail in Chapter 6.

The way in which data are organized will be determined to a very great extent by whether they are held on punched cards for mechanical processing or on a computer file. The purpose of data organization is to facilitate the storage and extraction of the data and their analysis.

Data coding

Data organization normally implies a coding process of some kind and whether the registry is manual or computerized, the basic principles are the same.

As far as possible registries should endeavour to use internationally recognized coding schemes. In the first place, these have usually been drawn up by a committee of experts, the combined wisdom of which will greatly exceed that available to a single registry, and the result is likely to be a better scheme. Secondly, adherence to international standards is the only sure way to achieve international comparability and the adoption of internationally agreed coding for the major data items is a self-evident advantage in making inter-registry comparisons. Recommended codes for various items are given in Chapters 6 and 7.

A registry is likely to need to develop its own coding schemes to deal with local data items, for example, to code its hospitals and consultants. It is a good idea to build into the coding system used for a data item some sort of structure, preferably one which has an element of classification where appropriate. As far as possible, the type of analysis or selection which will be required of the data item later on should be envisaged. There is the tendency on the part of some designers of coding schemes to compile a list of the terms to be coded, put them into alphabetical order and then apply a series of numerical codes.

It is worthwhile, especially if a computer is being used, to expand the codes beyond just their discriminatory function. Perhaps three digits are sufficient to identify all the consultants treating patients who are reported to one registry, but it may be worthwhile adding a fourth digit to the code to identify the consultant's speciality—for example, general surgeon, radiotherapist, gynaecologist etc. A tabulation presenting numbers of cancer referrals by speciality would be very simple if this coding scheme was adopted, while without it the analysis would be extremely awkward to specify.

When designing coding schemes it is important to examine the data item to be coded and to understand its nature. It should not be assumed that all variables can be classified, and hence coded, on one axis, i.e. in one dimension. Some data items have several dimensions, for example, diagnosis, which is recognized by the *International Classification of Diseases for Oncology* (ICD-O) as being essentially a three-dimensional variable—site, histology and behaviour. These are treated as if they were three independent variables and thus it is possible to code in any combination (see Chapter 7). Another example of a multi-dimensional data item is occupation. While many occupations are only pursued in one industry, for epidemiological work it may be important to know the specific industry in which, for example, a process-worker is employed. The most satisfactory way to deal with this at the coding level is to regard occupation and industry as a two-dimensional variable and design the scheme accordingly.

Data validation

It is very important to ensure that the quality of the data is as high as possible. This will be considered further in Chapter 9 but in a well designed system, particularly a computerized one, data validation is part of the data organization function. By definition invalid data cannot be organized correctly whereas incorrect data can. It is not possible to detect all incorrect data—for example, a patient may be reported to the registry as being born on 15 July 1923 whereas he was actually born in 1932, the year digits having been transposed. The data item is incorrect but valid and the error will probably be unnoticed unless the age is recorded and used to cross-check or another report is received which has the correct date of birth. A transposition of the day digits to 51 July 1923 is both incorrect and invalid and should never be allowed to be stored in the data-base. Systems should always be designed to detect invalid data, including invalid codes, as early as possible and this should be built into the data organization procedures of the registry.

Documentation of data organization

It is inevitable that, as a registry develops, changes to its data structure are made. New data items may be introduced and certainly it will be necessary to create new codes from time to time. All of these changes should be fully documented so that data users know what to expect from the data. Unfortunately, some changes have to be made almost on the spur of the moment to react to some new situation or as a result of an arbitrary decision about an individual case. All too often, instructions are given verbally or in the form of a memo on a single sheet of paper. Registries should have formal documentation giving the details of all changes to the structure of the database, including the date new codes were introduced or old ones discontinued.

Major coding revisions

With the passage of time some coding schemes need to undergo major revision. While a registry may be able to avoid this in its local schemes, international codes are revised periodically and the registry is obliged to follow. Careful consideration must be given to whether old records are to be converted to carry the new codes so as to preserve the continuity of the data, or if a clean break must be made at a certain point—preferably at an incidence year—and two (or more) consecutive schemes used.

The latter procedure should be followed only if data conversion is impossible—either because the data are processed manually or because the coding schemes do not allow for meaningful, accurate conversion. Discontinuous coding schemes are a major potential source of coding errors since almost certainly, both schemes will be in use together for a time as new registrations for cases belonging to the earlier period arrive together with cases for the later one. There can also be very serious difficulties arising in the analysis of such data and in the design of computer systems to maintain them. Whenever a new coding scheme is considered, every effort should be made to ensure that it is forward compatible from the old one. Data conversion should be identified as one of the factors to be taken into account when costing and planning the implementation of new coding schemes. If code-conversion is carried out, this must also be thoroughly documented because it is almost inevitable that this will subsequently affect the interpretation of the data.

Physical organization of manually processed data

Although many registries have held their data in the form of punched cards, which can be counted on mechanical sorters and tabulators, the introduction of electronic data-processing has rendered most of this machinery obsolete, and registries still using these methods would be strongly advised to become computerized as quickly as possible.

Edge-punched cards have been used in some registries but it is doubtful whether these have any realistic future. They can only be used for small numbers of patients—probably less than 1000 per year. If resources are really limited, it would be possible to hold data of this volume using a home computer costing less than US $1000.

If data are to be held entirely manually, it is traditional practice to maintain three physical files. These comprise the patient index file, arranged in alphabetical order as described earlier in this chapter, the accession register, and the data card, or tumour record, proper.

The accession register is simply a listing of the cases registered, arranged in order of their registration, i.e., by the accession number itself. This register is, in fact, used to assign the accession/registration number to all new patients. The accession register should include, as a minimum, the year of registration, the accession/registration number, the patient's name, and the primary site of the tumour.

The data card is the physical record containing details about each individual tumour which is registered. This may take several physical forms—as well as the registry abstract form, a variety of punch cards have been used in the past, as described above. Usually these tumour records are kept in numerical order within site, so that there will be a box of lung record cards, stomach record cards etc. This will make for easier counting, since any counts will almost certainly be by site category. There may also be some advantage in having cards of a different colour for males and females since counts are usually also made with respect to sex.

Physical organization of computerized data

This is an extremely complex subject since the options available are wide and the implications of each option are considerable. The matter is dealt with at some length in the *Directory of Computer Systems Used in Cancer Registries* (Menck & Parkin, 1986), and only a brief outline will be attempted here. The choice of medium is normally between magnetic tape and disk.

Magnetic tape storage

Magnetic tape files consist of a series of records, each cancer case probably occupying one record while each magnetic tape contains many thousands of records. Because files may spread over more than one reel, there is effectively no size limit to the file and, in applications outside cancer registration, files of many millions of records are not uncommon. The old restriction of punched cards which limited each record to eighty characters does not normally apply to magnetic tape records, though sometimes the programs used impose inconveniently small limits. Magnetic tape records are processed serially, that is, they are read or written in the order in which they are held on the tape. Normally, records on a magnetic tape are not altered *in situ*. If changes are required or new data are added, it is necessary to write an entirely new tape which contains the altered and new records as well as all the records which have not been changed. Records are deleted by simply not copying them from the old to the new tape. Because it may take over an hour to copy data from one tape to another, even on a large computer, it is obviously not possible to change one record at a time. Hence alterations are saved up and performed in batches—perhaps several thousand alterations are carried out on one run. This is known as batch processing and is a characteristic of magnetic tape systems. Because, as a form of storage, magnetic tape is relatively cheap, most registries still use this as their primary data medium. It is not, however, particularly convenient to carry out analysis of large tape files, since the records have to be ordered numerically while most analyses are oriented to a specific site, or group of sites. Some registries, therefore, have duplicate records which are arranged diagnostically in different files—one for lung cancers, another for stomach and so on in much the same way as recommended for manually held data.

Disk storage

The methods used for storing data on disk are rather more complex than those used for magnetic tape. One of the major advantages of disk over tape is that it is possible to process the records in any order, irrespective of their physical position on the disk. Thus the alteration of one record at a time is possible and, usually, the operator, using a VDU, can communicate directly with the data. A case can be displayed on the screen, altered and rewritten if necessary without disturbing any other records in the system. This is known as online processing, and it opens up many new possibilities for efficient use of the computer. Interactive record linkage has already been discussed and powerful coding techniques will be considered presently. As data can be processed in any order, the computer must 'know' where the record is physically located, even though the operator does not. This is achieved by the setting up of pointers in an index file which is maintained by the system. By means of carefully designed indexing techniques, data may be accessed randomly (as with an operator using a VDU) or in various indexed sequences—numerical, alphabetical, diagnostic and so on. The data are stored only once, and each of the indexes used has a pointer to every record.

While there are a number of excellent commercial software packages available to maintain data-bases of this complexity, considerable expertise is necessary in the detailed specification of systems using them, and the advice of computer professionals must be sought before embarking on the design of software of this nature.

Coding techniques

Manual coding

As has been indicated above, the main purpose of coding is to provide an organization of the data to allow efficient analysis. Manual coding is straightforward in that it consists of looking up the term to be coded in a coding manual and recording the code to be used. In fact, experience, training and skill is required because the terms used on the registration documents are not always given in the coding manual. Thus coding clerks using ICD-O would need to know that a tumour described as 'Intra-duct adenocarcinoma, invasive' is not coded as 8500/2 'Intraductal adenocarcinoma' but 8500/3 'Infiltrating duct carcinoma'.

Coding requires a great deal of concentration on the part of the coding clerks, as mistakes are easy to make, and, while some may be detected at a later stage, many will not. It is probably wise to set limits on the number of cases which are coded by each clerk each day, since tiredness may well give rise to unacceptably high error rates.

Those involved in the management of the registry carry a high level of responsibility for the accuracy of the coding. It is absolutely essential that sufficient coding manuals are available and that these are kept up-to-date and in good condition. Proper training must be given and adequate supervision provided. Rules must be well documented and any major changes carefully field-tested before introduction. Failure to think things through at the outset can result in frequent changes which are irritating for coding staff and inevitably lead to errors of one kind or another.

Computerized coding

The introduction of online processing has enabled some registries to reduce the amount of manual coding, or even eliminate it altogether.

The basic theory of computerized coding is exactly the same as that of manual techniques—i.e., looking up terms in a dictionary and extracting the appropriate code. In the case of computerized coding the contents of the coding book are stored on disk, the operator enters the text to be coded, usually using a VDU, and the computer searches among the texts in its dictionary to establish the code. Most systems use a method of preferred terms and synonyms. Each code used is associated with one preferred term and a variable number of synonyms. This is best illustrated by an example.

In the morphology section of ICD-O (WHO, 1976), the code 8070/3 is associated with the preferred term 'Squamous-cell carcinoma, NOS' (NOS, not otherwise specified) but other terms also appear so that the entry is given as follows:

8070/3	Squamous-cell carcinoma, NOS
	epidermoid carcinoma, NOS
	spinous-cell carcinoma
	squamous carcinoma
	squamous-cell epithelioma

The terms indented are all synonyms for 'Squamous-cell carcinoma, NOS' and all are associated with the code 8070/3. Thus, the operator may enter the term 'Epidermoid carcinoma NOS' and the computer generates the code 8070/3. When this data item is subsequently decoded, either for display on the terminal or as a print-out in an analysis, it would be translated to 'Squamous-cell carcinoma, NOS', its preferred term, the original text being lost. Terms other than those appearing in a coding manual may also be added, including any accepted abbreviations—almost certainly 'SCC' would appear in the example above. Alternative forms omitting the 'NOS' would also be entered as synonyms, as would the commonest misspellings of some terms. Computerized coding systems may also include procedures for editing texts before they are coded. This is useful for expanding abbreviations which may occur in various contexts—for example 'Ca' to 'Carcinoma', or to remove punctuation characters or redundant words such as 'Gland' if these have not been entered in the dictionary.

Of course, difficulties arise when terms which appear on cancer registration documents are not found in the dictionaries. This happens during manual coding but, whereas in the latter case the coder must select the most appropriate code to apply to the given term, the computer-coder must select and enter another term which is appropriate to the given text. This may, on rare occasions, mean referring to the coding book, but will be made more convenient by building into the system procedures for displaying the relevant part of the dictionary on the screen, from which the operator may select the most appropriate term. New synonyms may constantly be added to the dictionaries so that it is the computer which 'learns' rather than the operator. The degree of operator skill should not be underestimated, however. The coding of medical data often unavoidably involves a degree of interpretation and this requires an understanding of the terms used and experience in their use. Computerized coding undoubtedly increases the efficiency of the coding clerks and almost certainly enhances the accuracy of the coding. It does not necessarily mean that less training of the staff is required or that workers of inferior calibre can be employed. Skilled clerks are still required, but in smaller numbers.

Data dictionaries

It is appropriate to consider the use of data dictionaries here because it underlines the importance of relating the data organization at input to the data organization at output.

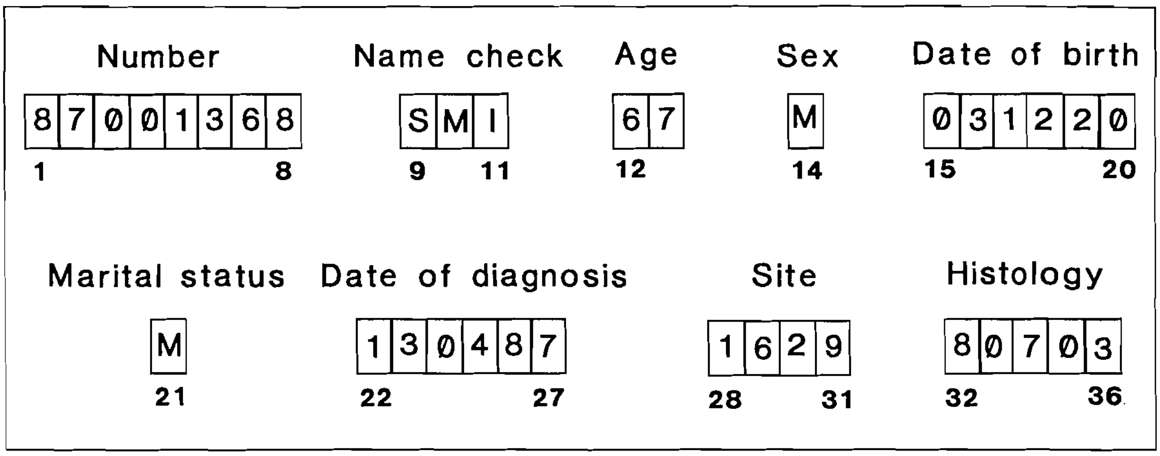
A data dictionary is a table that defines, for each data item, its name, where in the computer system it is stored, how it should be processed on entry and how it should be processed on output. It may also contain the specification of any validity checks that may be carried out on it and may specify under what conditions the item is present or absent. One great advantage of data organization through a data dictionary is that the program instructions are independent of the application, in other words one program may be used to drive many systems because the detailed specification is defined in the dictionary. This can be printed out to provide hard-copy documentation of the system. If modifications are required, it is the data dictionary which is changed and no actual programming is necessary.

In order to get information out of a system it is necessary to know how it was put in, and the data dictionary provides that information. Analysis software can be designed so that the user simply has to specify, for example, which variables are to be cross-tabulated, and the computer can find the location of the items within each record, perform the tabulation and, when printing the results, use as labels the terms corresponding to the codes encountered in the data dictionary. The data dictionary can also be used to document changes to the system—when items were introduced or discontinued, or coding systems were changed.

Medium conversion

When using computerized systems, it is necessary to present the data to the computer in a machine-readable format. Punched cards were frequently used for this, though their use has largely been superseded by key-to-tape or key-to-disk systems.

When data are manually coded, the coder must write the code into boxes printed either on special coding forms or incorporated on the source document itself. A typical completed coding form is shown in Figure 3. Each coding form must carry the identification number and there is also a name-check (in columns 9–11) to guard against amending the wrong record. It is very important to adopt conventions regarding the punching of certain characters—for example, to differentiate between zero and alphabetical O. Clear writing is essential to avoid ambiguity and punching errors and to maintain adequate punching speeds. When coding is done on the



Number	Name check	Age	Sex	Date of birth
8 7 Ø Ø 1 3 6 8	S M I	6 7	M	Ø 3 1 2 2 Ø
1 8	9 11	12	14	15 20

Marital status	Date of diagnosis	Site	Histology
M	1 3 Ø 4 8 7	1 6 2 9	8 Ø 7 Ø 3
21	22 27	28 31	32 36

Figure 3. A typical coding form

abstract form itself, the design of this document will allow for this. Some items can be self-coding, and can be entered directly from the form. Two obvious examples are:

Sex	1 Male	Marital status	1 Single/never married
	2 Female		2 Married
			3 Widowed
			4 Divorced
			5 Separated
			9 Unknown

Other examples can be derived from the suggested coding schemes for data items provided in Chapter 6. In these examples, the coder simply marks the appropriate category, and the data entry clerk enters the corresponding code. Self-coding minimizes coding and transcription errors, but only a limited number of items can be dealt with in this way. More complex variables must be coded into special coding boxes, which may appear in the margin of the form, or adjacent to the text of the item to be coded.

The coded forms are passed to a key-operator who types the codes, together with any textual or numerical data directly into a computer, or into a machine that produces either punched cards or records on tape or disk, which can be subsequently input to the computer. To avoid punching errors, each form may be typed again or verified, and any differences between the first and second attempts are indicated and checked to see which one is correct.

Where online systems are used and data are keyed directly into the computer via the VDU, medium conversion is not necessary. Any corrections are made there and then, and verification is usually unnecessary because a visual check is made at the time of entry.

When computer systems are designed, thought should be given to procedures for outputting data onto magnetic media for transmission to other registries or research organizations. The ability to pool comparable data is an important factor in many research applications and if this can be done using magnetic tape or floppy disks the amount of work required is greatly reduced. It is generally much more satisfactory to

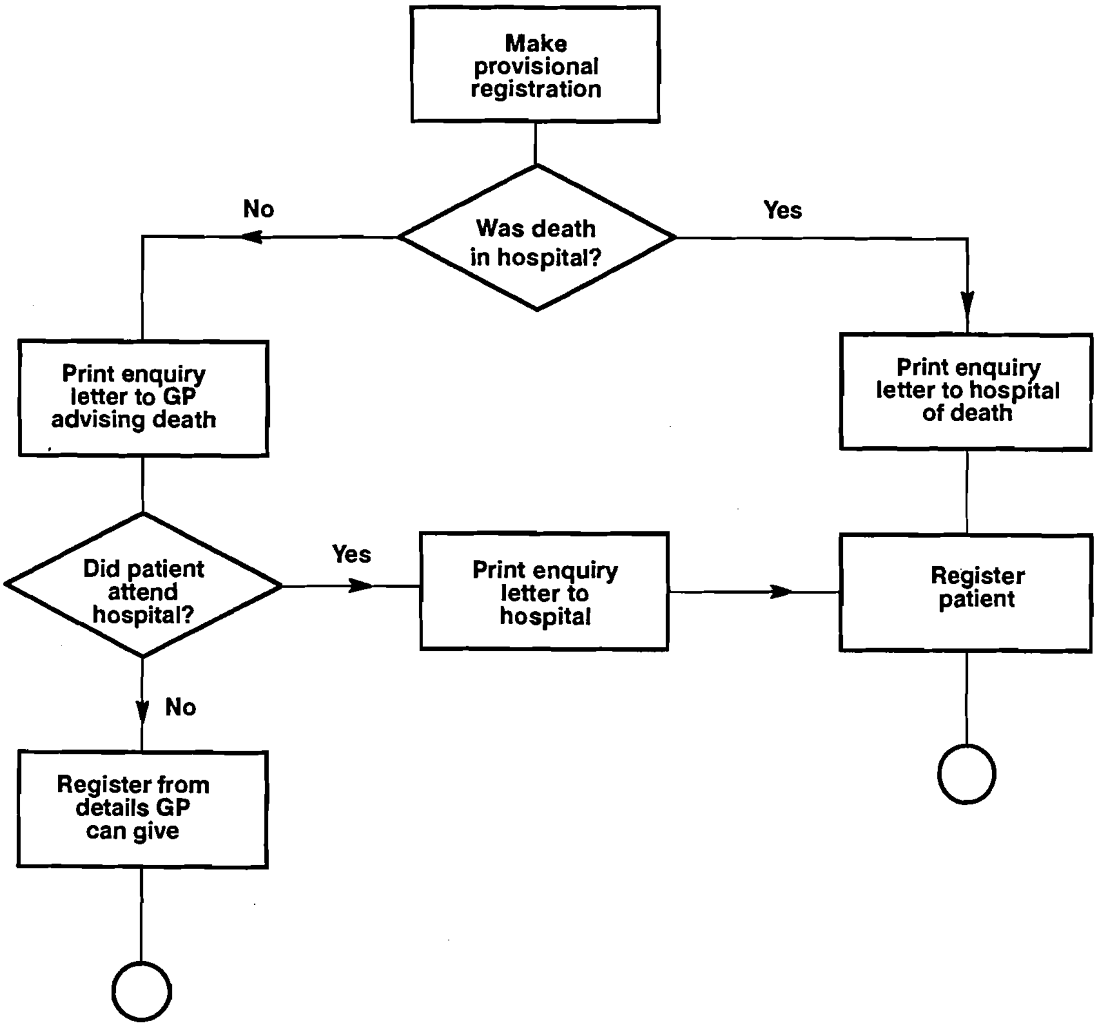

Make provisional registration
Was death in hospital?
No
Yes
Print enquiry letter to GP advising death
Print enquiry letter to hospital of death
Did patient attend hospital?
Yes
Print enquiry letter to hospital
Register patient
No
Register from details GP can give

Figure 4. Generation of enquiry from death certificate

pool data before carrying out a single analysis than to combine the results of many separate, albeit identical, analyses. Routines for extracting data onto magnetic tape should always be part of any computerized cancer registry system, though if the system is entirely disk-based, an intermediate computer will have to be used.

Enquiry generation and follow-up

Enquiry generation

In order to maintain high-quality data, registries frequently have to make additional enquiries, either to obtain complete data where these are missing or to resolve any inconsistencies which occur in the data already held (see Chapter 5). In the manual registry., this may take the form of sending out standard letters giving the patient's details and nature of the problem. A highly computerized registry may have in-built routines to automatically generate enquiries when the system itself detects that data are missing or inconsistent. These enquiries would be printed on a weekly or monthly basis. An example of how such a system may be designed to generate enquiries following the receipt of a death certificate for an unregistered case is shown in Figure 4. A provisional registration is made on the basis of the information on the certificate.

If the patient died in hospital an enquiry letter is sent (in the case of a computerized system, automatically generated) to the hospital in which the death occurred, and at the hospital the case is abstracted in the normal way. If the patient died at home or in a nursing home, an enquiry is sent to the doctor who certified the death, usually the patient's general practitioner, asking for details of any hospitalization the patient has had or, if there was none, for basic details of the diagnosis and, in particular, the date first seen for the disease. For hospitalized patients, a further enquiry is generated, this time to the hospital concerned to enable an abstract to be made. Such a procedure could be manual or computerized.

Where inconsistencies in information have occurred, these should be resolved at the data source, usually the hospital. In some cases there may be difficulties in determining the exact diagnosis at the registry. This commonly occurs in the case of a second tumour, which may be either a recurrence or a new primary, and the information given in the case notes is equivocal. Difficulties also occur when a registration is rejected by a computer because the site and histology appear to be inconsistent. In such cases, it is usually a good policy for the registry director to write personally to the clinician caring for the patient. This provides the registry with as good a solution to the problem as can be achieved, but serves also to directly remind the doctor that the registry exists and is prepared to go to some trouble to ensure that its data are as accurate as possible. It is important that these enquiries do not have the appearance of being mass-produced and are only made in cases of genuine difficulty. This represents an important component of the registry's task of continuously cultivating relationships and developing confidence. Such enquiries almost invariably yield further, unsolicited information about the same or similar cases subsequently.

Follow-up

Active follow-up

Registries operating an active follow-up system make enquiries, usually annually, about each patient thought to be alive. The enquiry is usually generated at around the anniversary of the first treatment. In manual systems, index cards of all patients still subject to follow-up are kept in boxes according to the month when follow-up is due, and forms are sent either to hospitals or to general practitioners as appropriate. The card for a patient is removed from the follow-up index if the patient is reported to be dead, either as a result of a returned follow-up form or when a death certificate is received.

For computerized registries using a batch system, the follow-up requests are automatically printed from the computer file in the appropriate month, and usually this is incorporated in the registry's update system. Online systems using active follow-up will have an index file based on the anniversary dates from which the requests will be printed. The registration or accession number, the patient's name and address and any other necessary details are transferred to preprinted forms using continuous stationery. The forms themselves are printed in addressee order to avoid manual sorting. Such systems are almost totally automatic, so few staff resources are required for their production.

Passive follow-up

Registries operating a passive follow-up system rely on external sources for the notification of all deaths of registered cases, irrespective of whether the death was due to cancer or to some other cause, and irrespective of where the death occurred. Thus no routine enquiries are generated but the information received in this way may give rise to *ad hoc* enquiries, for example, when the cause of death is given as being of a cancer other than one for which there is a registration.

Other important aspects of cancer registry operation

Two other matters should be considered, both concerned with the physical security of data.

Document control

It is important that all documents sent to a cancer registry are acted on appropriately. In large registries, the amount of paper present can be quite enormous, and it is essential for the maintenance of good data quality that information is not lost. Whenever forms are sent to or from the registry, counts should be made so that any losses can be identified quickly. Processed and unprocessed documents must be filed quite separately and this means that adequate storage facilities must be available. Clear policy decisions must be taken as to what source documents should be retained, for how long and in what form (microfilming may become necessary), and what documents may be safely destroyed after they have been processed. Arrangements must be made for the secure and confidential disposal of all documents which carry the names of patients if these are to be destroyed. Proper procedures should be adopted for the passing of information to other registries where this is appropriate.

Physical security of documents and computerized data

It is most important that as much protection as possible is afforded against the loss of both paper documents and computer files. This applies both for reasons of breaches of confidentiality and because of the value of the data itself. Equipment and buildings can be insured against loss or damage and these can be replaced. Replacement of documents and computer files is usually only a remote possibility and precautions must be made to ensure that the chance of loss is minimal. Paper documents can only be made secure by ensuring that they are stored under conditions which will guard against fire, flood and interference, since it is usually impractical to keep copies. Computer files, both programs and data, should always be kept at least in triplicate. Data on disks must be backed-up regularly so that in the event of hardware failure or accidental deletion, the data can be recovered.

At least one copy of the data should be securely stored away from the registry itself. When a major reorganization of computerized data becomes necessary, sufficient copies of the original data should be made so that, should anything go wrong, the original data can be reproduced. These copies should be retained indefinitely, since obscure but important errors in data conversion may not come to light until long after the conversion has taken place.

Computerized registries should have audit procedures, not only as part of the updating system but also as free-standing programs. These should be run at regular intervals to ensure that data are not inadvertently lost. This is particularly important for magnetic tape systems using multi-reel files where recovery from tape failures can sometimes result in cases being lost without being detected. Registries which are relatively minor users of large computer installations at remote sites are particularly vulnerable to accidental data loss. It seems to be a law of nature that the only computer files which get lost or become corrupted are those for which no copy is available!