Chapter 18
Designing, planning and conducting epidemiological research

18.1 Introduction

In previous chapters of this book, we covered the basic methodological principles of study design, analysis and interpretation. In this chapter, we concentrate on practical aspects of how to design, plan and conduct an epidemiological study. Only general issues are covered. Each research project and each study setting presents unique problems which cannot be dealt with here.

The first step in the design of any epidemiological study is to have an 'idea' for a study. This is a creative process for which no guidelines or advice can be given. The idea for a study usually comes from one's own work and experience and from the realization that there is a need to obtain a clear answer to a particular research question. Once the idea for the study has been identified, the next step is a critical review of the existing literature to find out what exactly is known about the subject and to make sure that the question has not already been answered. Where appropriate, registers of clinical trials and/or directories of on-going research (such as that of the International Agency for Research on Cancer) should be consulted. It may also be helpful to contact experts in the subject. In addition to revealing what is already known about the question, this will help the investigator to become familiar with problems that other researchers have faced, when using various study methods.

Once all the relevant information on the topic has been gathered, the next step is to state the study hypotheses in a clear and practical way. There is a tendency to formulate hypotheses that are broad and vague. Instead, they should be stated in terms of clear, simple, answerable questions. All main and secondary hypotheses should be formulated at the beginning of a study to ensure that all the necessary data are collected. They should include a concise description of the exposure(s) or intervention(s) to be studied, the outcome(s) of interest, and the magnitude of the anticipated effect(s) in the population in which the study will be conducted.

18.2 Preparing a protocol and obtaining funding

If we are convinced that the study is worth doing and that it is feasible, the next step is to obtain the necessary funding. If the total cost of the project is small, it may be possible to conduct it with available local resources, but many epidemiological studies are expensive enterprises which require substantial external financial support.

18.2.1 Funding bodies

Funding bodies which may be potentially interested in the proposed research should be identified. Some funding bodies do not fund cancer research. Others have their own research agenda and will fund only projects that address the particular questions in which they are interested. In some countries, directories of funding bodies with their areas of interest and funding budget are published regularly.

It is important to obtain grant application forms from all potential funding bodies and read the documentation carefully, checking the areas they are particularly interested in funding, and the deadlines for submission of applications. Sometimes, their material include lists of proposals they have funded in the past. These lists give an idea of the areas they are likely to fund in the future and the amount of money usually allocated to individual projects. It may be beneficial to contact funding bodies at this stage to check whether the proposal is of potential interest to them; if the proposed study is outside their areas of interest, there is no point in making the effort of preparing and submitting an application.

Writing a protocol

The protocol of the study should be written according to the specifications of the funding body. Although the layout of the application forms varies from one funding body to another, they are generally divided into the following sections:

Study hypotheses

The main and secondary study hypotheses should be written in a clear and concise way, indicating the exposure(s) and outcome(s) of interest and the magnitude of the anticipated effect(s).

Background and justification

This section should include a brief review of the state of knowledge about the topic. It should 'justify' the need for the proposed study by clearly indicating its originality and the potential significance of its findings. The proposed research may be a logical extension of previous work conducted by the researchers or of an initial pilot study. The results of such studies should be presented here.

Study population and methods of recruitment

The geographical location and the demographic characteristics of the study population should be described. Any particular reasons for the choice of the study population should also be given. For instance, a particular study population may have been chosen because of its exceptionally high exposure to a particular risk factor or for logistic reasons such as ease of follow-up. Details should be given on how the study subjects will be recruited.

Study design

It must be made explicit whether this will be an intervention, cohort, case–control, cross-sectional or routine-data-based study. The choice of design needs to be justified (see Chapters 5 and 7–11).

Sample size

It is necessary to show that the proposed number of subjects in the study will provide adequate power or precision to detect or estimate a particular effect. Sample size estimates should be presented under different assumptions. A single estimate is rarely convincing (see Chapter 15).

Methods of data collection

The methods of data collection (e.g., interview, laboratory measurements, extraction of data from clinical records, etc.) should be described in sufficient detail to show that the plan has been adequately prepared and is feasible. Possible practical constraints and strategies to overcome them should be presented here.

Statistical analysis

A concise description of the statistical methods planned for use in the analysis of data should be given (see Chapters 6 and 14).

Ethics

The protocol of any study involving human subjects should provide answers to the following questions: What will the subjects be told? What will their collaboration entail? Will invasive procedures be used? Are there any risks for participants? How will consent be obtained (e.g., at an individual or community level; written or verbal)? What steps will be taken to ensure confidentiality of the data? It will be worth consulting at this stage the ethical guidelines proposed by the Council for International Organizations of Medical Sciences (CIOMS) and the World Health Organization (WHO) (1993). Most funding bodies will fund only projects that have been approved by the relevant ethical committees and some will wish to see samples of the information sheet and consent form to be given to the study subjects.

Timetable

A realistic timetable for carrying out the various activities of the study should be provided. This should incorporate milestones to be achieved at regular time intervals (e.g., every six months), which will help in monitoring the progress of the study.

References

References to key publications should be included in the grant proposal.

Budget

The budget is generally divided into staff costs, equipment (e.g., computers, freezers, centrifuges) and running expenses (e.g., office expenses, computer and laboratory consumables, travel costs). Each item must be justified and of reasonable cost. This will often involve discussions with the personnel office about staff costs and consulting price lists to get the best prices for supplies and equipment. In certain countries, allowances should be made for inflation and currency fluctuations when calculating the final budget.

It is also important to check whether the institution where the study is going to be based will charge 'overheads'. These correspond to the costs to the institution of administering the grant, organizing the payment of salaries, ordering supplies, providing office space, heat, electricity, air-conditioning etc. Some funding bodies refuse payment of overheads (e.g., the World Health Organization), whereas others will pay only up to a certain proportion of the total cost (e.g., the European Commission pays only 20% of the total cost). If the funding body does not pay overheads, it may be possible to include as running costs in the budget some expenses that might otherwise be covered by the overheads, such as telephone, fax and mailing costs.

Dissemination of results

Some funding bodies require applicants to state how they intend to inform study participants about the findings of the study and how they will be disseminated to relevant health authorities and the scientific community. This topic is further discussed in Section 18.4.

Box 18.1 presents a checklist on how to prepare grant applications. Before sending the grant application to the funding body, it is useful to send a draft to people who have experience of writing and reviewing applications, for comment. Although no one can foresee all the problems that may occur, many potential difficulties will be quickly spotted by experienced researchers. In particular, it is advisable to involve a statistician in the development of the study protocol.

The investigator should also inform all people whose approval or cooperation is either required or desirable. Proposed research in clinical or academic institutions should be presented to appropriate departmental heads and/or hospital administrators. Often there will be a committee specially designated to review and approve study protocols. Studies conducted in the community should be presented to local health officials. In addition to gaining the required approval, the investigator may receive valuable practical suggestions and other assistance from these individuals, such as introductions to physicians who may permit the study of their own patients. The investigator may also learn of other similar or related research that is under way. Cooperation with other investigators may help avoid duplication of effort and lead to sharing of resources and, possibly, even of data.

Box 18.1. Checklist on how to prepare grant applications

- Identify funding bodies potentially interested in the proposed research. Obtain grant application forms from them.
- Read their guidelines carefully. Check deadline for submission of applications.
- Check whether they require submission of any special documents (e.g., statement of willingness to participate in the study from all collaborators, confirmation of ethical approval) which may take some time to obtain.
- Make a list of the staff and equipment required. Consult personnel office to obtain estimates of staff salaries. Obtain list of prices for equipment. Make an estimate of running expenses (e.g., office expenses, computer consumables, travelling expenses). Check whether your organization charges overheads.
- Write the study protocol according to the instructions of the funding body, respecting the various section headings. Do not exceed the space and/or maximum number of words specified for each of the sections.
- Be concise and clear. Check spelling errors carefully. If you are writing the application in a language other than your mother tongue, it will be useful to show it to someone who is proficient in that language.
- Show the draft of the protocol to experts in the topic. Review the draft in the light of their comments.
- Make sure you send the original application and the requested number of copies to the funding body in time to reach its office before the closing date for submission of applications.

Review of grant proposals

Most funding bodies send the applications to experts in the particular subject. These experts are asked to evaluate the proposals by answering the following questions: Does the proposal address an important question? Is the study design appropriate? Are the applicants qualified to conduct research? Is the budget reasonable? Each referee will send a report with his/her comments to the funding body.

Most funding bodies have their own research review panel who will consider the applications and the referee's reports. The panel will rank the applications and award those on the top of the ranking scale. Occasionally, applicants may be invited to submit a revised application which takes into account the referees' comments. The revised application should address all the referee's comments and be submitted together with a covering letter detailing the revisions.

18.3 Conducting the study

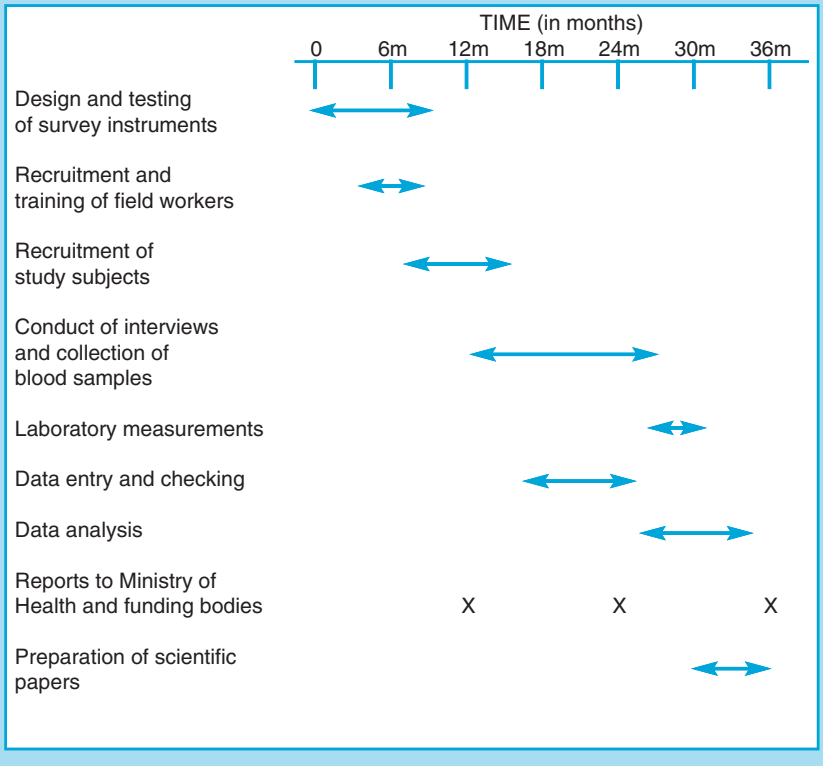
Once funding for the study is secured, the researchers need to start planning the conduct of the study carefully and in detail.

18.3.1 Logistic organization

It is important to plan all the activities of the study at a very early stage. Three general principles should govern the logistic organization of a study. First, realistic milestones should be set up at the start of the study. Second, all study procedures and decisions should be properly documented. Third, expenses should be monitored closely.

A *timetable* of all the activities to be conducted should be established. An example is given in Box 18.2. The timetable should indicate targets to be achieved at regular time intervals. It must be realistic. For instance, it should take into account possible delays in delivering vital equipment, climatic factors which may affect the fieldwork, and staff holidays and sickness.

Box 18.2. Example of an organizational timetable



A *study manual* should be prepared in which each study procedure is documented in detail (e.g., step-by-step instructions for the administration and completion of questionnaires or for the collection of biological samples). The manual should be updated if any changes are made during

the study. Copies of the relevant parts of the manual should be given to all research team members. In addition to the study manual, it is also important to keep a *study diary* in which all problems encountered are noted and the solutions adopted are recorded. This will be useful in maintaining consistency of decisions throughout the study.

A detailed *record of expenditure* should be kept to monitor expenses throughout the study and to facilitate submission of expenditure statements to the funding body. It is important to keep copies of salary payment sheets, invoices, receipts, order forms, etc. Grant accounts can be easily monitored by using a computer spreadsheet.

18.3.2 Recruitment, training and management of personnel

Epidemiological studies often require the recruitment of personnel (e.g., statisticians, interviewers, data-entry clerks, computer programmers, etc.). A job description should be written for each post indicating the responsibilities involved, and the minimum qualifications and experience required.

It is important to establish a clear line of management so that each staff member knows what his/her responsibilities are and to whom he/she should report. Proper training of personnel is fundamental to the success of the study and adequate time and resources should be allocated to this. Performance of each staff member should be monitored regularly and constructive feedback given. Successful work must be seen as a team effort.

18.3.3 Equipment

The equipment required will depend on the design of the study and the setting where it will be carried out. Routine-data-based studies may require only access to good computing facilities. In contrast, a large intervention trial conducted in a remote area may require everything from computer equipment to staff accommodation and transport facilities.

Computing equipment

The choice of computer equipment ('hardware') and computer programs ('software') (see Box 18.3) will be determined by the design and size of the study as well as by the availability of technical expertise for data-processing and statistical analysis and of local servicing facilities. However, the rate of development of computer equipment and software is such that any advice or guidelines soon become out-of-date. Thus, it is essential that professional advice, from a computing specialist, should be sought at the planning stage of the study so that advantage is taken of the most recent developments.

Hardware

Computers differ in their type of microprocessor, the size of their random access memory (RAM) and their storage capacity. The most commonly used microcomputers are based on the 486 and Pentium micro-

processors. The RAM governs the size of program and the amount of data that the machine can actively work with at any time. The bigger the RAM the better. Most machines have at least 4 Mb of RAM, but if Windows is to be used, at least 16 Mb is desirable. The data storage capacity is determined by the capacity of the hard disk. In general, the hard disk capacity should be at least 500 megabytes (Mb), especially if Windows-based software is to be used, as these programmes tend to use a considerable amount of space on the hard disk.

Once the appropriate type of machine for the study has been selected, it is necessary to obtain price lists from different manufacturers. Machines with similar specifications are produced by many different manufacturers and the prices tend to differ widely. However, value for money is not the same as cheapness. Equipment reliability, warranty and free maintenance from the dealer are all important factors which need to be considered. The availability of good local servicing and repair facilities is another important consideration, particularly for studies in developing countries.

Microcomputers with similar features are available in two basic models: desktop/tower and laptop (portable) computers. Laptop computers have the advantage of being portable, but they are more expensive than desktop machines of equivalent computing capacity. However, since they work from rechargeable batteries, they avoid the need for an uninterruptible power supply (see below).

It is absolutely vital to take adequate precautions against the possibility of losing data that have been entered onto the computer. All the information on the computer should be regularly copied onto some other storage device to guard against the possibility of hard disk or computer failure. Floppy disks are adequate to store relatively small data-sets, but removable hard-disk units may be needed for very large data-sets. Alternatively, a 'tape-streamer' may be used to back-up the data onto a small magnetic tape. It is essential to make back-up copies of the data regularly, on both a daily and weekly basis, and to create multiple copies as a back-up to the back-up (see Section 18.3.6).

Unreliable power supplies can lead to loss of data and serious damage to the computing equipment. It is advisable to avoid sharing the mains circuit with equipment such as electric motors and air-conditioning systems that makes heavy but intermittent demands on the power supply. Power stabilizers (or mains filters) are small devices, sometimes incorporated into the plugs, which are designed to deal with rapid voltage fluctuations. However, they offer no protection against loss of data if there is a power failure. Uninterruptible power supplies (UPS) are devices that, in the event of a power failure, provide power (from batteries) for short periods, until the batteries run down. This gives enough time to save the work and switch off the computer without loss of information. Some UPSs shut down the computer automatically should the loss of the main power supply be prolonged.

Printers vary in price enormously. In general, the better the print quality, the higher the price. Dot-matrix printers are relatively cheap and the quality of their output is usually good enough for most requirements. An ink-jet or a laser printer may be required if there is a need to produce higher-quality documents such as grant proposals, questionnaires and study reports.

Box 18.3. Computing requirements

- Hardware (machines)
  - *Minimum requirements*
    - computer
    - printer
  - *Optional requirements*
    - back-up facilities (e.g., tape-streamer) or removable hard drive
    - uninterruptible power supply (UPS)
- Software (computer programs)
  - *Minimum requirements*
    - operating system
    - database package
    - statistical package
    - word-processing package
    - virus-checking software
  - *Optional*
    - data transfer software

Computing equipment can be used in most normal ambient conditions. It is advisable, however, to try to ensure that they are kept in a dust-free environment with temperatures below 25°C (to avoid overheating) and humidity of less than 80%. Thus, special equipment to ensure that ambient conditions are appropriate (e.g., air-conditioning systems) may be required, particularly in hot and humid locations.

Software

In order for a computer to work, it requires a set of instructions which control its operations. This set of instructions is called an operating system. In the past, most IBM-type microcomputers used the DOS operating system, but today the Windows operating system is almost universal. Windows allows several programs to be run simultaneously (which is not possible with DOS unless a special software package has been installed) and, therefore, it is easier to move data, text and figures between different programs.

Apart from the operating system, most epidemiological studies require a database package to enter and check the data, a statistical package to per-

form the statistical analyses, and a word-processing package for use in preparing the data-collection forms, progress reports and papers with the final results for publication. Many suitable computer packages are available. The best policy is to use well known packages that have been thoroughly tested by others; newly developed packages often contain 'bugs' (i.e., errors) that can cause major problems. Epi Info (version 6) is a very useful package for word-processing, questionnaire design, data entry and validation, and simple tabulations and statistical analysis. It has been developed by the Centers for Disease Control in the USA and the Global Programme on AIDS at the World Health Organization in Geneva. This software package was specifically developed for epidemiological studies, it is easy to use and it is distributed free of charge (apart from postage)[a].

Computers may become infected by a 'virus', which is a piece of software, written by unscrupulous people, designed to damage the computer or the data it holds. Precautions should be taken to prevent this happening and to detect any possible contamination immediately. Virus-checking software packages have been developed to routinely check for the introduction of viruses and to clean them from the computer. Floppy disks should be scanned for viruses before data are transferred into the computer. It is important to ensure that the program is updated every time new viruses appear.

If there is frequently a need to transfer data between computers (for instance, from a portable computer to a desktop computer), this can be done efficiently via a cable link which connects the serial or parallel ports of the two machines. Special software will be needed for such data transfer. Data can be transferred much more quickly this way than by using floppy disks. If constant access to the data is required from more than one computer, a network of computers can be created. However, this is more complex to set up and specialist help should be obtained.

Other equipment

It may be necessary to obtain office equipment such as desks, filing cabinets, a photocopying machine, a fax machine, etc. Studies involving collection of biological specimens may require laboratory equipment such as freezers, centrifuges, etc. (see Section 18.3.5). Field studies in remote areas may require staff accommodation and transport facilities.

18.3.4. Recruiting study subjects

The recruitment of study subjects may involve obtaining lists of potentially eligible subjects. Such lists may be obtained from local authorities (e.g., lists of all subjects resident in a particular area), from doctors or hospitals (e.g., lists of patients with a particular condition) or from cancer registries. If a list of all the residents of a particular geographical area is required as a sampling frame and no enumeration lists are available or those available are grossly inaccurate or out-of-date, it will be necessary to conduct a census of the defined study population before recruiting the study subjects.

[a] Details can be obtained from USD Incorporated (2075A West Park Place, Stone Mountain, GA 30087, USA; telephone +1 404 469 4098).

The recruitment of subjects usually involves writing letters and/or meeting the study subjects, their doctors, local health representatives, leaders of the communities, etc. to explain the objectives of the study and to obtain their support and consent.

Once recruited into the study, each subject should be given a *unique study identification number*. This number should remain constant throughout the study. It will help to keep confidentiality and blindness and will be used to link individual information obtained from different sources (e.g., questionnaires, clinical records and laboratory forms).

18.3.5 Collection of biological samples and laboratory measurements

Many cancer-related epidemiological studies involve collection of biological specimens such as blood, serum, plasma, urine, cell and tissue samples from study participants.

Collection and processing of biological specimens

Biological specimens may be specifically collected for the purpose of the epidemiological study or may derive from samples collected for clinical purposes. Clinical activities can be an easy source of biological materials, but the specimens are sometimes inadequate for an epidemiological investigation, particularly if collected by a number of different people using different methods of collection and processing. The amount of specimen available may be insufficient in volume or quality, specimens may have remained too long in less than optimal conditions of preservation or they may be contaminated with extraneous materials (see Box 18.4).

Blood can be collected by venepuncture or by finger (or heel) pricks. Finger (or heel) pricks tend to be more acceptable to study participants. The volume of blood obtained with this method, although smaller than that obtained by venepuncture, is usually sufficient for most laboratory assays. It is, therefore, important to check with laboratory staff the minimum amount of blood that needs to be collected to perform all the necessary laboratory assays. The timing of the blood collection may also be important (e.g., time since exposure, time of the day, time within the menstrual cycle, etc.). Some laboratory assays require fasting samples. Other factors such as a subject's posture and the use of a tourniquet can affect the concentration of certain blood components.

After collection, blood can be separated into several components including serum, plasma, red cells and white cells. Separation must be done shortly after the blood has been collected (ideally, within two hours). During this period, the specimens can be left at room temperature, but preferably at 4°C. Serum and plasma samples should be frozen as soon as possible after separation, and stored at –20°C. This is adequate for most assays, at least for several weeks. Some assays require immediate storage at –70°C or below (e.g., vitamin C measurements).

Some urinary assays require collection of 24-hour urine samples, whereas others can be performed in a single sample collected 'on the spot'. The collection of 24-hour urine samples is logistically more complex and less readily

accepted by study subjects. Particular care is necessary in instructing study subjects in the collection of 24-hour urine, and in checking that instructions have been followed. It is usually necessary to leave the container with the individual overnight and to arrange for it to be collected the following day. Creatinine levels should be measured to take into account and correct determinations for possible losses of urine. A good check for the completeness of urine collection is the administration of 250 mg of *para*-aminobenzoic acid, which is completely recoverable in the urine over the 24 hours following its administration.

Before storing samples for long periods of time, it may be necessary to add appropriate fixatives and stabilizers. This should be checked with the laboratory staff. Biological samples are easily damaged by repeated freezing and thawing. Thus, samples should be divided into small portions before freezing. Ideally, the size of the aliquots should be such that each one contains just sufficient material to perform the assays required at a particular time.

All biological specimens should be properly labelled. To ensure confidentiality, the label should contain the unique study identification number of the subject but not any other personal details (e.g., name). The identification number will make it possible to link back the results from the laboratory assays to the records of the individual from whom the sample was taken. In some circumstances, it will be appropriate to include on the label the date of collection and the type of specimen, if not evident. It is important, however, to ensure that laboratory staff are kept 'blind' to the exposure (or outcome) status of the subjects from whom the samples were taken. Pre-printed, adhesive labels with each identification number duplicated several times are available commercially. Alternatively, they can easily be produced by a microcomputer. Waterproof marker pens should be used, except if samples are going to be frozen in salt–alcohol mixtures, in which case plain pencils should be used instead. The labelling must be done on the body of the container and never on the cap only. Numbers and letters must be written in a clear and standardized form (see Box 18.5). It is important to pre-test the labels to check how they stick and how pen/pencil writing is preserved during transport and storage.

Samples may be stored temporarily in a laboratory located where the fieldwork is being carried out before being sent to the laboratory where the assays will be conducted and/or to their permanent storage place. If a large number of samples is collected and stored, it is important to store them in a way that allows rapid retrieval of any particular sample. For instance, samples may be stored in batches according to the date they were collected or frozen. This information may be computerized so that any particular specimen can be easily and rapidly located.

Quality control of laboratory procedures

Laboratories should keep a record of the reagents, test kits, laboratory equipment, the batches of supplies and reagents used at different times, and the number of times each aliquot has been thawed and re-frozen.

Box 18.4. Collection, processing, transport and storage of biological samples

- Step-by-step instructions should be given in the study manual on how to collect, process, transport and store biological specimens.
- All the methods for collection, processing, storage and transport of specimens should be pilot-tested before being implemented.
- All biological samples should be considered potentially infectious. All personnel involved in their collection, processing, analysis or storage should take measures to protect themselves against the possibility of becoming infected (e.g., by wearing disposable gloves).
- The quality of laboratory measurements should be closely monitored. This consists of:

  Internal quality control
  – Use in-built standard controls.
  – Perform duplicate measurements.
  – Send a random sample of specimens twice, laboratory staff being unaware that they are duplicates.

  External quality control
  – Send random duplicate samples to a reference laboratory.

Records should also be kept of reasons for considering a specimen as 'unacceptable' (e.g., insufficient material, inadequate processing, storage or transport) and of unusual events which may affect the results of a test (e.g., power failures, errors in test procedures).

Many laboratory tests have 'in-built controls' using standardized reagents of known concentration or quantity. In addition, many assays are performed in duplicate as a normal routine laboratory procedure. Although in-built controls and duplicate measurements are important to assess the quality of laboratory measurements, they are not sufficient (see Box 18.4). A random sample of specimens should be sent to the laboratory and tested twice, laboratory staff being unaware that they are duplicate samples. The reliability between the first and the second measurements gives an indication of the internal quality of the laboratory procedures (see Section 2.6.2). Ideally, reliability should be checked within batches, between batches and from one time period to another (e.g., week-to-week). It is important to measure intra-observer (by having duplicate samples processed by the same observer at different times), inter-observer (by having the same samples processed independently by two different staff members) and inter-product reliability (by analysing the same samples with different batches of reagents). High reliability does not ensure high

validity of the measurements, however. Thus, it is desirable to send a random sample of duplicate specimens to an external reference laboratory, whose measurements will be taken as a 'gold standard' (see Section 2.6.1).

Laboratory results should be recorded in laboratory notebooks or, preferably, on specially prepared forms to facilitate computer entry. It is important to ensure that the forms are designed in a way that allows the recording of particular problems or features which may be of relevance to the interpretation of the measurements. For example, it should include items on the identity of the technicians involved with each test (to check for inter-observer variation), the batch of reagents used (to check for inter-product variability) and any technical problems that may arise (e.g., lost and broken samples, samples with insufficient material, errors in test procedures, etc.).

18.3.6 Data processing and editing

All the various steps of data processing and editing should be planned early at the design stage of the study. This should be done in consultation with a statistician and a computer programmer.

Design of data-collection forms

All forms to be used for recording data in a study should be carefully designed and pre-tested to ensure that the relevant data are collected and can be easily extracted for data processing. This general principle applies to all data-collection forms: questionnaires, laboratory forms, data extraction forms from hospital notes, etc. These issues were discussed in Chapter 2 and Appendix 2.1.

Data collection and recording

Adequate training is the single most important aspect of data quality control. It should focus on accuracy and completeness, with an emphasis on how and why the work of the entire project is dependent upon the quality of the data recorded. Field workers should be instructed how to write numbers and letters so that the coding and data entry clerks have no difficulty in reading the forms. Some letters and numbers are frequently confused and special instructions should be given regarding these. Examples are given in Box 18.5.

The leaving of blanks should be avoided since these can be interpreted in different ways. A blank may mean that the question was not appropriate, the answer was not known or, simply, that the question was mistakenly skipped by the interviewer. In addition, some computer programs interpret blanks as zeros, which may be a valid code. It is preferable to design the forms so that all options are covered (for example, the answers to a specific question may be: 'yes', 'no', 'not known' and 'not appropriate').

Information should be recorded in pencil or using a ball-point pen. Errors should be corrected by writing the correct response above or below

Box 18.5. Handwritten numbers and letters which are commonly confused and possible solutions to avoid confusion

Characters confused	Solution
1 and 7	This confusion arises if ones are written with an initial upward stroke (for example 1). If this is the situation, always write sevens with a horizontal line through them (as do the French), that is 7. A simpler solution may be to insist that ones are written with a single stroke (for example /).
O (oh) and 0 (zero)	Note 1. Write 0 (zero) with a line through them, that is Ø
4 and 9	Write 4 as 4
4 and 7	These digits may be confused if sevens are written with a horizontal line through them! Instruct field workers to ensure that the top of the seven is written horizontally
6 and 9 (upside down)	Relevant when coding laboratory samples (for example, is it 6I or I9?) Draw a horizontal line under all numbers, for example 19
2 and Z	Note 1. Always write Z with a horizontal line through it, that is Ƶ
5 and S	Note 1. Always write 5 using two pen strokes
O (oh) and Q	Note 1
I and 1	Note 1. Always write I with 'hat and shoes', not as a single stroke, that is I not | .
U and V	Avoid both letters as codes as far as possible (but they will be needed for names)

Note 1. Avoid using the alphabetical character in data fields that may contain alphabetical or numerical information.

[a] Reproduced with permission from Smith & Morrow, 1996)

the original response but never writing over the top of the incorrect response. Field workers and supervisors may use pens of different colours so that it is clear where corrections were made.

The data-collection activities should be supervised and the quality of the data monitored on a regular basis. Every data-collection activity should be designed in such a way that it can be checked. Checks should be designed with the objective of detecting errors rather than proving the high quality of the data.

Forms should be checked for their data completeness and consistency (Box 18.6). Interviewers' performance should be monitored regularly and constructive feedback given. If feasible, a random sample of subjects should be re-interviewed by the interviewer, co-worker and/or supervisor to assess the reliability of the measurements.

Interim tabulations and scatter plots of the most important variables should be produced regularly. These may help to detect problems. For instance, systematic patterns in the data (e.g., one interviewer with a much higher refusal level than the others) will indicate the need to check on particular aspects of data collection.

Box 18.6. Ensuring good data quality

- Ensure adequate training and supervision of field workers.
- Check samples of data-collection forms to assess their completeness and accuracy.
- Assess interviewer's performance by watching/listening to interview.
- Re-interview a random sample of subjects. The second interview may be conducted by the supervisor, by another interviewer, or by the same interviewer. Assess reliability of the data obtained in the two interviews.
- Tabulate the most important variables by interviewer to assess inter-interviewer variability.

Coding

Coding is the term used to describe the conversion of data from a data-collection form into a format that is suitable for analysis on a computer, by assigning a numerical code to every possible answer on the data-collection form. For instance, sex may be coded 1 for males and 2 for females; only the number 1 or 2 will be entered onto the computer file. Numerical data (e.g., number of children) do not require coding since the exact number can be entered.

Many data collection forms are 'pre-coded', that is, they are designed so that every possible answer is assigned a code in the form (see Appendix 2.1). The field worker selects one (or more) of these 'pre-coded' options. Such pre-coded forms have the advantages of being almost ready for computer entry by the data clerk and of minimizing transcription errors. However, later coding is still required for 'open' questions and for answers which do not fit into the pre-coded categories. These answers may be given additional codes if necessary.

An alternative approach is to code the data only after the form has been completed. The data-collection forms are designed so that there is a spe-

cific column for coding on the right- hand side of the page. The field workers (or the study subjects if the questionnaire is self-administered) are asked to ignore this column and fill in the answers in the spaces provided on the left-hand side. The answers are coded later by the field workers or by specially trained coding clerks and the codes are inserted in the right-hand column. The time between data recording and coding should be kept as short as possible so that inconsistencies in the data are revealed and the field worker can be asked to re-contact the subject to correct them.

Some aspects deserve special attention when developing coding instructions. Firstly, numerical variables such as number of children, weight or height should not be coded into pre-determined categories. Their actual values should be recorded so as to preserve the full detail of the data in analyses. Data may then be grouped in different ways, if appropriate. Secondly, specific codes should be given if a particular question is not applicable to a respondent (e.g., number of pregnancies for a man) or the answer is not known or the response was mistakenly left blank by the field worker (for example, 1 = yes; 2 = no; 7 = not recorded; 8 = not relevant; 9 = not known). Similarly, specific result codes should be developed in laboratory forms to allow for the coding of technical problems such as broken samples, insufficient material, error in laboratory procedures, etc. To minimize coding errors, it is advisable, if possible, that the meaning of any particular code remains constant throughout the data-collection form (e.g., a code of 9 should be given to the 'not known' answer of every question). Thirdly, a coding manual with all the coding values and rules should be compiled and given to field workers, data-coding and data-entry clerks. This manual should be updated if any changes in coding are made during the study.

Computer data entry

Once the data are coded, the information on the forms is ready to be entered into a computer. For each type of data-collection form, a computer file should be created to enter the data. It is possible to set up a data-entry programme so that the computer screen resembles the layout of the data collection form to minimize errors by data-entry clerks. The data-entry program is usually developed by using a database package. Most of these programs incorporate consistency and range checks so that the data are checked and edited in an interactive way. For instance, if an inadmissible value is entered, an error message appears on the screen and an audible warning alerts the data-entry clerk to the error. It is also important to design the data-entry program so that it does not allow data to be entered from records which were erroneously given the same identification number.

Each variable for which data are recorded must be given a name for use in computer processing. Variables may be identified by numbers (e.g., 'VAR001'), but it is better to use abbreviated names that easily identify the variables, such as 'sex', 'age', 'parity', 'fbirthag' (for age at first birth). The

number of characters to be used in a variable name should not exceed the number allowed by the computer program (most packages do not allow the use of more than eight characters).

If data relating to the same individual may be obtained from different sources (e.g., interviewer-administered questionnaire, clinical notes and laboratory reports) or follow-up surveys conducted at different points in time, it is better to store the information from each source or survey in a separate file. It is relatively easy to link data from different database files using a common person-identifying code (i.e., the unique study number).

Each data-set collected in a study should be entered onto the computer twice. The second entry should be independent of the first and should be done by a different data-entry clerk. The two files of data should be compared using a specially written computer program (available in Epi Info). Any differences detected by this program should be checked against the original records and the incorrect file(s) edited. The program to compare the files should be run again and the editing process repeated until no differences between the files are found.

A manual with the names of the files where data from each form have been entered and with details of all the variables (what the variable is, meaning of codes, allowed values, etc.) should be compiled and distributed to all data-processing personnel. It should be updated on a regular basis, as necessary.

Data cleaning

Range and consistency checks should be run for each variable, either while they are being entered into the computer or after a whole batch has been entered. Range checks identify inadmissible values. For instance, if the variable sex was coded '1' for males and '2' for females, values other than 1 and 2 would be flagged as errors. Consistency checks cross-check the consistency of data for related variables. For example, males should not have a diagnosis of ovarian cancer.

In addition to checking for incorrect or unusual values, the distribution of each variable should be examined. Interim tabulations of the data and scatter plots of quantitative variables, such as height against weight, are effective in revealing unlikely outlying observations. These should be checked against the data collection forms, since they are unlikely but not necessarily impossible. In some cases, it may be possible to correct the data. In other cases, it will be necessary to insert a 'missing value' code, if it is certain that the data were in error (e.g., an impossible height).

Updating and storage

Back-up copies of the data should be made regularly (ideally on a daily basis, but if this is not possible, at least weekly). A minimum of two back-up copies of all data on computer should always exist. Some of the copies should be stored in a geographically separate location in a dry and relatively dust-free environment to guard against events such as fires, floods, robbery, etc. A

complete and updated list should be made of what data are on all stored disks and tapes, and copies should be kept at more than one location.

The data collection forms should also be stored in a secure place to ensure confidentiality of the data. The forms should be filed in batches or in serial order (either by date of collection or by study identification number) to facilitate their retrieval for checking apparent errors detected in records on computer. Ideally, the forms should be kept well after the end of the study, but space restrictions may preclude this. In large studies, it may be worth considering the possibility of copying the data collection forms onto microfiches.

Preparing data for analysis

The variables entered into the computer files are not always the ones suitable for data analysis. Recoding and computing of new variables is likely to be necessary. For example, body mass index (BMI) may be calculated from height and weight, and age at the time of the survey from date of birth and date the interview was completed. Moreover, continuous variables are usually grouped for the 'classical' methods of analysis based on stratification. Some combination of groups may also be necessary for categorical variables with large numbers of categories or for categories in which there are no, or very few, cases.

It is preferable to keep a copy of the original file (as entered from the data-collection forms) and create new files for the new versions of the data. Preferably, the names of consecutive versions of the same file should be selected so that it is easy to recognize which one contains the latest version (e.g., STUD1, STUD2, STUD3, etc.). It is important to keep copies of the programs written to recode variables and compute new ones as a document of what exactly was done and to allow new versions of the data to be re-created from the original file if errors are detected.

Careful thought should be given to the planning of the analysis of the data (see Chapters 3, 4, 6 and 14). The steps given in Box 14.1 should be used as a general guideline.

18.4 Disseminating the results

Research projects are worthwhile only if their findings are properly disseminated to the scientific community. This usually involves publication of the findings in peer-reviewed scientific journals. Results may also be presented at scientific conferences and meetings. Depending on the aims and relevance of the study, it may also be appropriate to produce more detailed reports and/or reports written in lay language to be distributed to health authorities, community leaders, study subjects, etc.

18.4.1 Writing a paper

The process of writing a scientific paper has many similarities to that of writing a grant proposal. Thus, most of the issues presented in Box 18.1 are equally relevant to the preparation of scientific papers.

The first step is to identify journals that are likely to be interested in the topic of the study and are appropriate for the importance of the findings.

Box 18.7 lists some of the most important international journals of relevance to cancer epidemiologists. A useful way of choosing an appropriate journal is to see where most previous work on the same topic was published.

The journal's requirements—style, headings, referencing system, etc. must be followed closely. 'Instructions for Authors' are generally published in most issues of journals. The draft of the paper should be prepared exactly according to instructions. It is important to think carefully about how to illustrate the results (tables, figures, etc.) so that the main findings of the study are clearly presented to potential readers.

Most journals ask the authors to structure the paper into the following sections:

Title page

This page should include an informative and concise title for the paper; the name(s) of the author(s); the name of the institution in which the work was performed; the institutional affiliation of each author; and the name and address of the author to whom correspondence should be addressed.

Abstract

The abstract should summarize in no more than a few hundred words the reason for the study, the major findings, and the principal conclusions. Some papers (e.g., *British Medical Journal*) require a structured abstract which is divided into special sections. It is helpful to obtain some examples of abstracts published in the journal before attempting to write one. Some journals ask the authors to provide key words from the Medical Subject Headings of *Index Medicus* for indexing purposes.

Introduction

This section should include a description of the background of the investigation (citing only essential references) and the reason(s) for performing the study.

Data and methods

The methods (including the statistical approach) used should be described in enough detail that someone reading the paper could repeat the study. For research on human subjects, it is necessary to highlight any ethical issues that the conduct of the study might have raised, whether approval was obtained from the relevant ethical committees and how consent was obtained from the study participants.

Results

The findings should be presented in a clear and concise way. Repetition should be avoided by presenting data in only one format, that is, results displayed in a table should not be presented in the text.

Discussion

The discussion should cover the issues presented in Chapter 13: strengths and limitations of the study design, quality of the data, bias, confounding and chance. Finally, the findings of the study should be discussed in the light of previous work.

Acknowledgements

Contributions, including technical help or other assistance, and funding bodies that provided financial support for the study should be acknowledged.

References

The references should be written according to the specifications of the journal where the paper is going to be submitted.

Tables and figures

General rules for the preparation of tables and graphs were given in Section 3.4. Journals may have specific rules, which should be followed carefully.

Before submitting the paper, it is important to send a draft to all collaborators, as well as to colleagues who are successful at publishing articles and ask for their comments. If needed, the manuscript should be revised in the light of their comments before sending it to the journal

18.4.2 Review

If the editor thinks the journal might accept the paper, he/she will send it out to referees. The reports from these referees will then determine whether the paper is accepted or rejected. Papers are rarely accepted without alterations. Most commonly, it will be either rejected or accepted subject to modifications based on the referees' comments. The authors should submit a revised version of the paper which takes into account these comments, together with a covering letter to the editor explaining exactly how this was done.

Box 18.7. Some international journals of potential relevance to cancer epidemiologists

American Journal of Epidemiology
British Journal of Cancer
British Medical Journal
Cancer
Cancer Causes and Control
Cancer Research
Epidemiology
European Journal of Cancer
International Journal of Cancer
International Journal of Epidemiology
Journal of Epidemiology and Community Health
Journal of the National Cancer Institute
Lancet
Statistics in Medicine

Further reading

* The book by Smith & Morrow (1996) provides useful practical information on how to plan and conduct epidemiological studies. Although the book focuses on field trials, most of the issues presented are relevant to any type of epidemiological study.