RESEARCH DESIGN

Data collection and management for research studies

Ruth Morley

All research studies, be they randomised clinical trials or epidemiological or observational studies, require appropriate, accurate, and complete data.

The most important step in research (not covered in this paper), is study design. Unless there has been careful definition of the hypothesis to be tested, outcomes to be measured, and calculation of required sample size, time and resources will be wasted.^{1–4} Researchers should consult a statistician or epidemiologist for help with this crucial stage.

There are already a number of excellent articles on the design of data collection forms.^{5–7} This paper is based on 12 years of personal experience of designing and using data recording systems for randomised trials (including thousands of children^{8,9}) and observational studies and is applicable to studies conducted by research personnel or health professionals. Design of questionnaires for completion by patients or parents (covered in a recent article¹⁰) is not discussed here.

Important stages in the data collection process

- (1) Defining data required for the study.
 - Outcome measures
 - Identification, background, and preintervention data
 - Information on known or possible factors influencing outcomes under consideration
- (2) Designing and testing questions to obtain the required data.
- (3) Designing and piloting data collection forms or computer data input programmes.
- (4) Coding responses.
- (5) Assembling the data file and checking and cleaning data.

deafness in a population) and in a study in which index and control populations are compared an accurate definition of the index condition will be required.

Identification, background, and baseline data

A confidential record should be kept of each subject's identification. The requirement will vary from study to study but items to be considered are:

- Subject's study number
- Full name
- Sex
- Date of birth
- Caregiver's name and title
- Address (including postcode)
- Telephone number
- Contact (in case family move)
- Family doctor/paediatrician
- Date of data collection

Background and baseline data will need to be considered carefully for each study and a decision made about what is relevant and necessary. For example, if the study were of infant growth, birth anthropometry, gestation, and parental anthropometry would be among essential items.

Factors influencing outcomes under investigation

In a randomised trial outcome measures can be heavily confounded by factors unrelated to the intervention being studied and it is wise to be able to demonstrate balance for these factors (and determinants of these factors), in the randomisation. An example would be the influence of social class and parental IQ or education on a child's IQ. Known and possible factors should be identified from published studies and from the advice of experts. In an epidemiological study testing a causal hypothesis it is important to collect data not only on the explanatory variable of primary interest (for example exposure to radiation as a cause of leukaemia) but also on independent factors associated with either the disease or explanatory factor (which could confound or augment the effect of the explanatory factor). Decisions about which variable is of primary interest and which other data are to be collected should be based on what is already known about the disease, explanatory factors, and associations.

Infant and Child Nutrition Group, MRC Dunn Nutrition Unit and Department of Paediatrics, University of Cambridge

Correspondence to:
Dr Ruth Morley, MRC
Dunn Nutrition Unit,
Downhams Lane, Milton
Road, Cambridge CB4 1XJ.

(1) DEFINING DATA REQUIRED

Outcome measures

These will have been defined in the study protocol at the design stage. In a randomised trial it is hypothesised that these factors will be influenced by the intervention. In an observational study these are the factors of primary interest (for example, the incidence of

(A) Triceps skinfold thickness in mm .

Birth weight lb oz

(B) Birth weight lb oz

Figure 1 Showing appropriate (A) and inappropriate boxes (B).

(2) DESIGNING AND TESTING QUESTIONS

Designing questions which are clear, unambiguous, and to the point is more difficult than might be supposed. It is essential first to decide (i) what *exactly* do I want to know? and (ii) how do I plan to analyse the data?

Questions like these should be avoided: (a) Was he late sitting, walking *or* talking? (yes or no); (b) Was he late sitting, walking *and* talking? (yes or no); (c) Was she *not* eating well? (yes or no); and (d) Is she eating *less well* than she used to? (yes or no). From answer (a) it will be impossible to differentiate between an otherwise normal child who walked late versus a child with global developmental delay, whereas answer (b) would not yield information on specific developmental problems. It would be much better to ask the age at which the child acquired each skill, individually, with specific definitions of skills. Examples (c) and (d) are difficult to answer as 'yes' or 'no'.

It is always desirable to break data collection down to the smallest and least ambiguous units and to make sure questions are easily understood and, whenever possible, objective.

Thought should be given to questions specifically designed to check for reporting or recording errors. For example, in one of our studies assessors needed to calculate each child's age in order to calculate IQ. Software used for data entry was programmed to calculate age from dates of birth and assessment. If the computer calculated age differed from the assessor calculated age this was flagged and IQ recalculated before entry.

(3) DESIGNING AND PILOTING DATA COLLECTION FORMS OR COMPUTER DATA INPUT PROGRAMMES

There are three main ways to collect data: (a) record data on forms for subsequent manual entry into a computer; (b) enter data directly into a computer; and (c) record data on forms for optical mark reading.

(a) Data collection forms

- Keep the form clear and neat. Do not use vertical typeface, consistent upper case text, or an unusual or small font; these all make reading the form more difficult.
- Arrange data collection spaces or boxes down the right hand side of the page⁵ for ease of spotting missing data and subsequent typing of data into a computer data file.
- Include instructions for dealing with mistakes, for example, *draw a horizontal line through the box and write the correct answer in the margin beside the box, with your signature.*

- Include subject identification on each page, you may drop a pile of forms!
- Arrange questions in the order in which it is natural to ask them, with history before examination or measurements.
- Make spaces or boxes appropriate for the answer. Text is recorded and read more accurately on an open line than in boxes.¹¹ Number or letter codes can be recorded in boxes, which should be appropriate for the expected range of responses, with the decimal point in the right place and units clearly stated either in the text or after the boxes, as in fig 1A. If the boxes are laid out as in fig 1B, a weight of 10lb 1oz, for example, could be filled in as illustrated and there could be confusion as to whether this represents 10lb 1oz or 10lb 10oz. This potential confusion is avoided with the boxes as illustrated in example (A).
- A 'don't know' or 'not sure' option may be needed. A decision should not be forced when there may be doubt (for example, does the child have a squint?).
- Have a code for missing data (making sure it is not used for anything else). This makes it easier to check for accidentally missing data. At the foot of every page we have the instruction:

Code ☒ if an item cannot be completed and record reason

- *It is essential to pilot forms* on subjects similar to the study population, ideally by observing form completion, debriefing the form filler, and examining completed forms. Revision and repiloting is usually necessary.
- Store completed forms in a secure place and *always keep a copy* of any forms sent for data entry or to a statistician.

(b) Direct entry of data into a computer

This option is increasingly available, particularly in the clinical setting.¹² There are pros and cons for collecting research data in this way.

Pros:

- Data entry software can be programmed to accept only data falling within predefined categories or range, minimising recording error.
- Programs can be written to check for inconsistencies or to calculate the likely range of one variable based on that of others, so unlikely values can be flagged and rechecked immediately.
- Data are entered once only, reducing workload and (if checked) the potential for error.

Cons:

- Unless data are regularly backed up and/or printed out there is a risk of loss from power or system failure.
- Typing data into a computer takes slightly longer than filling in a form and disrupts the flow of questions and answers and rapport with the subject or parent.
- It can be more difficult to record volunteered (out of question order) information than on a form, where answers can

(A) Hearing test result	(test Omitted, Normal result, Retest arranged, Specialist referral, hearing Loss)	[N]
(B) Hearing test result	(Test omitted=1, normal result=2, retest arranged=3, specialist referral=4, hearing loss=5)	[2]

Figure 2 Showing two main coding possibilities: alliteration (A) and numeric (B).

be recorded in a different section with ease.

These systems should be piloted in the same way as data collection forms.

(c) Forms for optical mark reading

This method is not covered here and is not favoured by the author because of constraints on form layout, problems with reading accuracy, and the need for special equipment and dedicated personnel. Researchers wishing to use these should seek technical advice.

(4) CODING OF RESPONSES

Accuracy of data recording depends to a large extent on good question and form or screen design, but the way in which multiple choice questions are answered may also affect accuracy.

Data collection forms or screens can be designed so that: one of several options is circled or (on screen) pointed to with a touch pen; one of several labelled boxes is filled in (by a tick or cross); or a coded answer is written or typed into a box.

Barnard *et al* have shown that 'nothing is as easy for form fillers to do as simply ticking alongside the appropriate action'.¹³ However, like Gore,⁵ the author favours coded answers in boxes, largely because both checking forms for missing data and data entry are easier. It is relatively simple to spot an empty box but less easy to spot that none of a number of options has been circled or marked. With direct computer entry, however, data input software can check for missing data and further data entry is unnecessary, so there would be no reason to favour coded answers in boxes.

There are two main coding possibilities: alliteration and numeric coding and these are illustrated in fig 2 (completed for a child with a normal hearing test result). There are no reported studies to help decide which method generates the most accurate data and choice is likely to be decided on the basis of personal preference.

Alliteration is more intuitive and works well when there are few possibilities (for example, M or F for male or female). However, it becomes more complicated and less intuitive if there are multiple responses, though two letter codes can be used, as in the excerpt from a

Types of medication prescribed for the child

AB	Antibiotics	LI	Linctus
AG	Analgesics	LX	Laxatives
AS	Antispasmodics	ST	Steroids
BD	Bronchodilators	VI	Vitamins
HP	Hypnotics	OT	Other
IR	Iron	ZZ	None

longer list of codings shown in the table. Many statistics programmes require data to be numerically coded for some analyses, so recoding may be necessary. This is time consuming and errors are possible during the recoding process.

Numeric coding is the best option for coding ordered categories, like degree of severity of a symptom or sign (with the most severe category assigned the highest value). Furthermore, numeric data will not need to be recoded, eliminating work and a possible source of error.

Dates can cause problems. It is essential to find out how the chosen statistics software handles them and enter them in the right format. Many statistics packages internally convert dates to day counts, so elapsed time between two dates (for example, birth and measurement), can be calculated. If the software can handle dates entered as *day/month/year* they should be recorded in this format; it is the most intuitive and errors are least likely.

When coding (or recoding letter codes into) numeric data *the presence of a condition or giving of a drug or carrying out an intervention should always be given a higher value than its absence*. For example, it is recorded whether a child was given a drug that could influence the number of days in hospital. In regression analyses the direction and size of a regression coefficient will represent the effect of going from a lower to a higher value. Thus, if 'not given' is assigned the value 1 and 'given' 2, a regression coefficient of *minus* 0.5 for this factor means that children given the drug spent 0.5 *fewer* days in hospital. If 'given' were 1 and 'not given' 2, the coefficient would have been *plus* 0.5 (the influence of not being given the drug). The researcher would have to remember to reverse the sign to see the true influence of being given the drug, making interpretation of results unnecessarily complicated and confusing.

(5) DATA FILE ASSEMBLY AND CHECKING AND CLEANING OF DATA

If data have been recorded on a form they will need to be entered into a computer. Again a customised input program is desirable, preferably one requiring data to be typed in twice for verification (by comparing the two sets of data and flagging any discrepancies).

It is usually a relatively simple matter to read a data file into the preferred statistics package, but the inexperienced should seek advice. Even if checks were done at an earlier stage data should still be checked finally for outliers and inconsistencies.

If recoding is necessary this should also be checked for errors by keeping a copy of the original and recoded files and running a few simple analyses on both data files. For example, If M(ale) is recoded as 1 and F(emale) as 2, there should be the same number of subjects with the values M and F in one file as there are with values 1 and 2 in the new file.

Conclusion

Designing questions and a system for collecting appropriate, accurate, and complete data requires a surprisingly large investment of time, resources, and expertise but is vital for the success of a research study. The inexperienced researcher should seek advice from a statistician, epidemiologist, informatics expert, or colleague with experience of data collection and management.

The temptation to leave the details of data collection to the last minute and least experienced personnel should be strongly resisted.

The author thanks Tim Cole, Terence Dwyer, Colin Morley, and Patricia Wright for their helpful suggestions.

- 1 Altman DG. Study design. *Statistics in question*. London: BMA Publications, 1982: 3–5.
- 2 Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley, 1983.
- 3 Breslow NE, Day NE. *Statistical methods in cancer research*.

Volume II – The design and analysis of cohort studies. Oxford: Oxford University Press/IARC, 1987.

- 4 Schlesselman JJ. *Case-control studies. Design, conduct, analysis*. Oxford: Oxford University Press, 1982.
- 5 Gore SM. Assessing clinical trials – record sheets. *Statistics in question*. London: BMA Publications, 1982: 57–8.
- 6 Wright P, Haybittle J. Design of forms for clinical trials. *BMJ* 1979; ii: 529–30, 650–1.
- 7 Wright P, Lawrence G. Design of case report forms for clinical trials. In: Lloyd J, Raven A, eds. *ACRPI handbook of clinical research*. Edinburgh: Churchill Livingstone, 1994: 209–35.
- 8 Lucas A, Morley R, Cole TJ, et al. Early diet in preterm babies and developmental status at 18 months. *Lancet* 1990; 335: 1477–81.
- 9 Lucas A, Morley R, Cole TJ, Gore SM. A randomised multicentre trial of human milk versus formula and later development in preterm infants. *Arch Dis Child* 1994; 70: F141–6.
- 10 Fallowfield L. Questionnaire design. *Arch Dis Child* 1995; 72: 76–9.
- 11 Barnard P, Wright P, Wilcox P. The effects of spatial constraints on the legibility of handwritten alphanumeric codes. *Ergonomics* 1978; 21: 73–8.
- 12 Wyatt JC. Clinical data systems, part 2: components and techniques. *Lancet* 1994; 344: 1609–14.
- 13 Barnard P, Wright P, Wilcox P. Effects of response instructions and question style on the ease of completing forms. *Journal of Occupational Psychology* 1979; 52: 209–26.