Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Using databases in medical education research: AMEE Guide No. 77

Jennifer Cleland, Neil Scott, Kirsten Harrild & Mandy Moffat

To cite this article: Jennifer Cleland, Neil Scott, Kirsten Harrild & Mandy Moffat (2013) Using databases in medical education research: AMEE Guide No. 77, Medical Teacher, 35:5, e1103-e1122, DOI: 10.3109/0142159X.2013.785632

To link to this article: http://dx.doi.org/10.3109/0142159X.2013.785632

2013; 35: e1103–e1122 MEDICAL TEACHER

WEB PAPER

Using databases in medical education research: AMEE Guide No. 77

JENNIFER CLELAND, NEIL SCOTT, KIRSTEN HARRILD & MANDY MOFFAT
University of Aberdeen, UK

Abstract

This AMEE Guide offers an introduction to the use of databases in medical education research. It is intended for those who are contemplating conducting research in medical education but are new to the field. The Guide is structured around the process of planning your research so that data collection, management and analysis are appropriate for the research question. Throughout we consider contextual possibilities and constraints to educational research using databases, such as the resources available, and provide concrete examples of medical education research to illustrate many points. The first section of the Guide explains the difference between different types of data and classifying data, and addresses the rationale for research using databases in medical education. We explain the difference between qualitative research and qualitative data, the difference between categorical and quantitative data, and the difference types of data which fall into these categories. The Guide reviews the strengths and weaknesses of qualitative and quantitative research. The next section is structured around how to work with quantitative and qualitative databases and provides guidance on the many practicalities of setting up a database. This includes how to organise your database, including anonymising data and coding, as well as preparing and describing your data so it is ready for analysis. The critical matter of the ethics of using databases in medical educational research, including using routinely collected data versus data collected for research purposes, and issues of confidentiality, is discussed. Core to the Guide is drawing out the similarities and differences in working with different types of data and different types of databases. Future AMEE Guides in the research series will address statistical analysis of data in more detail.

Introduction

We are surrounded by data (facts) wherever we go. In our lives we constantly take in data from our environment, interpret and make sense of this and store relevant pieces of information for the future. Too much data can, however, lead to information overload and there is a limit to how much useful information an individual can effectively store and retrieve. This means that data need to be recorded more permanently for future reference. This is not new. Five thousand years ago as more complex societies developed, the need for accurate bureaucratic records led to an organised system of records on clay tablets. Other societies recorded information using stone, papyrus, paper or even knotted string and many employed scribes to record and interpret important data.

Such stored information is only useful if organised in such a way that it can be retrieved quickly. A database is just an organised collection of data for storing, managing and retrieving information. The term originates from the development of computing in the 1960s, but a database does not necessarily need to be in digital form. A filing cabinet or card box to store records in alphabetical order could be considered a database, as it is simply a means of storing data and is designed to enable fast access to such information (see Picture 1 for an example of a paper-based database).

An electronic database can be even more powerful than a paper-based database – not only can it store large amount of data, it can also sort and order data in convenient ways and establish connections and patterns between related records. For example, using a paper database you cannot quickly sort records by age, find the oldest or youngest person or find everyone born on a particular day. Using an electronic

Practice points

- A database is a tool for storing, managing and retrieving data, so it can be interpreted and used for various (in this case, for medical education research) purposes.
- Databases can be used to manage and analyse numerical and word-based data from quantitative and qualitative research projects.
- Deciding the nature of your research, understanding the nature of your data and how to classify data types are crucial to setting up your database.
- Practical considerations such as cost, available resources and support usually need to be taken into consideration, as well as the data management and analysis requirements of the project.
- Be aware of the local ethics requirements and international guidance for carrying out medical education research, and adhere to these. Being able to show how you addressed any issues of risk will help if you want to publish your work in a journal.

Correspondence: Professor J. Cleland, Division of Medical and Dental Education, University of Aberdeen, Foresterhill, Aberdeen AB25 2AZ, UK. Tel: 01224 437257; fax: 01224 437285; email: jen.cleland@abdn.ac.uk

ISSN 0142–159X print/ISSN 1466–187X online/13/051103–20 © 2013 Informa UK Ltd.
DOI: 10.3109/0142159X.2013.785632

Picture 1. A paper-based database.

database this information is readily available at the touch of a key.

Databases abound in everyday life now. When you go online to look for a cheap flight, the branded systems you use are sophisticated databases, organised in such a way that you can find the information you need. Similarly, online supermarket shopping is commonplace – when you go to the site of your favourite retailer to select the items you wish to buy and pay for them, you are using a database. When you search social media such as 'Facebook' ™ to find an old friend, you are again using a database. These public access databases are well planned to make them as 'user-friendly' as possible.

Organisations and companies, from small to large, heavily depend on databases for operations such as payroll, contact details for their employees and suppliers, ordering materials and so on. You probably also depend on your bank's database to keep track of your money and financial dealings.

The aims of this Guide are multiple. First, the focus is on setting up and using databases for medical education research purposes. We then introduce the different types of data – quantitative and categorical – and how to classify them. Related to this, we explain the difference between qualitative research and qualitative data. We present different types of database and data management software, for managing quantitative and categorical data. We then introduce how to work with quantitative and qualitative databases and some of the many practicalities of setting up a database so that it is usable for research. This includes: cases and variables, displaying data, variable names and value labels, unique identifiers, data entry, form design, dealing with missing data and 'handy hints'. Statistical analysis of data is beyond the scope of this Guide, but we provide guidance on how to organise your database and prepare your data so it is ready for analysis (such as 'eyeballing' and describing your data). The ethics of using databases in medical educational research, including using routinely collected data versus data collected for research purposes, are then discussed. Furthermore, the purpose of this guide is not to examine technical aspects of a database such as design, construction and maintenance, as these may require specialist programming skills. Rather, we focus on how to best use databases for medical educational research projects, such as identifying patterns of performance, which students struggle with communication skills, where your students go after graduation, and who interacts most with students on the wards and so on, and use examples to illustrate various points.

Different types of data

When dealing with data and databases, it is fundamental to understand the differences between the different types of data. Understanding the type or classification of your data is important for both how you enter it into a database and how you then analyse the data. One crucial basic point to consider is the difference between data and information. Data are a representation of information – words, numbers, dates, images, sounds, etc., without context. For example, here is a list of data items:

Fail
MCQ
100
Part 2
Pre-clinical
60

Data items need to be part of a structure, such as in a sentence, in order to give them meaning. Information is a collection of words, numbers, dates, images, sounds, etc., put into context to give them meaning. While you will probably

Table 1. The difference between information and data.

Paddy	Yes	
Susan	No	No
Rama	Yes	
Angelo	Yes	
By adding headings, the data becomes information		
Student Name	Attended ward round	Absence pre-authorised
Paddy	Yes	
Susan	No	No
Rama	Yes	
Angelo	Yes	

have spotted that these data relate to assessment in medicine, they do not gain true meaning until used in a sentence:

*The Year 1 (**pre-clinical** year) **MCQ** examination has **100** questions in its **Part Two**, of which students must pass **60** or they **fail** the course and must repeat the year.*

In other words, data can be thought of as raw material, while information is data that have been processed in such a way as to be meaningful. Databases are not, however, written in sentences to help you process the information they contain. Rather you need to use a structure in order for data to become information. In Table 1 the second and third columns contain either 'Yes' or 'No', but without headings there is no meaning.

This information, held in the database, can be used to determine which students are prone to not attending ward rounds without pre-authorising their absence. This knowledge then highlights which students need to be contacted by Faculty to discuss the reasons for their poor attendance at ward rounds. Knowledge is the ability to understand information and to then form judgements and opinions and to make decisions based on that understanding.

Classifying data types

Understanding how to classify data types is essential for both data analysis and for setting up a database. However, it can be a confusing topic as a number of different terminologies are used. Perhaps the most useful distinction that can be made is to divide data into two broad categories: quantitative or categorical ('qualitative'). Note that quantitative and 'qualitative' refer here to types of data, not to different types of research paradigm (see later). Just to confuse matters even further, quantitative data may also be referred to as numeric or scale data, while categorical data may be called qualitative or attribute data.

We will refer to qualitative data as categorical data from now on, to limit potential confusion between qualitative/categorical data and qualitative research (see later).

Categorical data is the measurement expressed not in terms of numbers on a linear scale, but rather by means of natural language description (e.g. eye colour = blue, height = short). A set of data is said to be categorical if the values or observations belonging to it can be sorted according to category. Each value is chosen from a set of mutually-exclusive categories (i.e. a subject can be in one group only). For example, people have the characteristic of 'gender' with categories 'male' and 'female', and they can be either 'male' or 'female', but not both. Similarly, people can be a member of just one age group or attend just one medical school.

Table 2. Different types of categorical data.

Nominal (no inherent order in categories)	Eye colour, ethnicity, diagnosis
Ordinal (categories have inherent order)	Job grade, age groups, course (e.g. Year 1 course, Year 2 course)
Binary (where there are only two categories)	Gender, grade for course (e.g. pass/fail)

Categorical data comprise either *ordered* or *unordered* categories. Unordered categories are where there is no natural ordering of categories, so they are nominally labelled (hence why unordered categorical data is also referred to as *nominal* data). Examples of unordered category data include marital status (e.g. single, married, widowed or separated), or different sports (e.g. baseball, basketball, football).

Ordered categorical data, or *ordinal* data, are defined by categories with a specific rank or ordering. For example, a question about student satisfaction with the following response options will collect ordinal data: unsatisfied, expectations met, exceeded expectations and significantly exceeded expectations. Categorical variables that judge size (small, medium, large, etc.) and attitudes (strongly disagree, disagree, neutral, agree, strongly agree) are also ordinal variables. A common example of ordinal categorical data is an asymmetric Likert scale, often used in questionnaire surveys. For example, you might be asked if you send emails to the student population every day, once a week, once a month, once a year, or never. (Note that the distances between points are not equal in this example, but symmetric Likert scales, with equal distances between points, are also commonly used).

Categorical data also encompass a third type of data: binary data – where there are only two categories (see Table 2 for examples of the different types of categorical data).

Quantitative data, on the other hand are numerical data. As this implies, this is measurement in terms of numbers. Usually, quantitative or numerical data are associated with a scale measure where the distance between categories is the same (e.g., the difference between 1 and 2 is the same as the difference between 2 and 3), unlike categorical data. Furthermore, quantitative data can be *discrete* or *continuous*. *Discrete* data are whole numbers (e.g. the number of complete years a student has been studying). *Continuous* data can take any value within a certain range (e.g. height, age, blood pressure). Discrete data have finite values – you can count them (e.g. number of days spent in hospital). Continuous data technically have an infinite number of steps, which form a continuum. The number of questions correct in a test is discrete, as there are a finite and countable number of questions. On the other hand, the time taken to complete a task is continuous, since time forms an interval from 0 to infinity.

See Figure 1 for a pictorial representation of these relationships between quantitative and categorical data and Table 3 for examples.

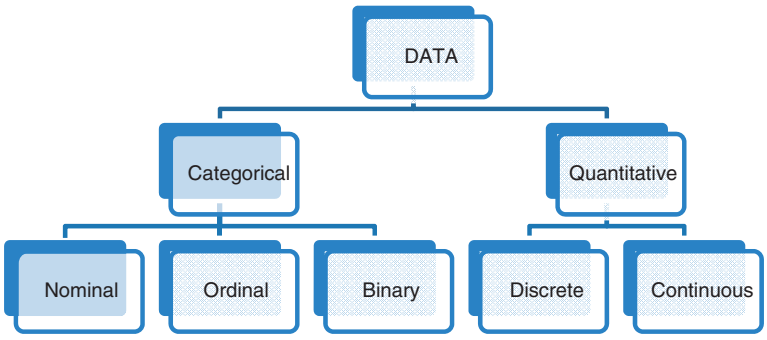


Figure 1. Different types of data.

Table 3. Examples of quantitative data and categorical data.

Quantitative data	Categorical data
Height	Gender
Weight	Religion
Income	Method of treatment
University size	Type of teaching approach
Group size	Marital status
Self-efficacy test score	Qualifications
Percentage of lectures attended	Native language
Clinical skills performance	Type of instruction
Number of errors	Problem solving strategy used
Age	Social class

As another example, we can describe what we know about JC (one of the authors).

Categorical She has green eyes (nominal data).
She is in her 40 s (ordinal data).
She is female (binary data).

Quantitative She has two legs (discrete data).
She has one brother and two sisters (discrete data)
She weighs 60 kg (continuous data).
She is 165 cm tall (continuous data)

Sometimes the distinction between data types is less clear. For example, age as measured using exact years could be considered either as quantitative (specifically continuous) or categorical (specifically ordinal) data. It all depends on the range of data and the context – when considering people of all ages, it would normally be treated as a continuous variable, but for a large group, you might be more likely to think of age as an ordinal variable (e.g. age range 0–10 years, 11–20 years and so on). In medical education, grades are commonly presented as ordinal (e.g. A, B, C, D, E), binary (e.g. pass, fail) or discrete (e.g. as a percentage) variables. Similarly, employment status is often thought of as an ordinal variable (when taking on values 'unemployed', 'part-time' and 'full-time'), but can be thought of as a nominal variable (when other values such as 'retired' or 'volunteer' are possible responses). A judgement call is often required when classifying data, based on the extent and amount of data.

Discriminating between qualitative research and qualitative data

The examples of 'qualitative' or categorical data given above might seem a little confusing – surely 'qualitative' data is text, graphics, document or visual data from interviews, observations, recordings, focus groups and so on? What is the relationship between qualitative research and 'qualitative'/categorical data? It is important to outline the difference here.

Fundamentally, what distinguishes the data in a qualitatively-designed study from 'qualitative' or categorical data is the set of assumptions and principles underlying the research (see Box 1 for an overview of research paradigms).

In short, quantitative and qualitative research methods are based on different paradigms which make different assumptions about the world (Guba 1978), about how science should be conducted (e.g. because of the type of research questions under investigation in qualitative research studies, sample sizes are usually quite small, but selected because they have certain characteristics – whereas quantitative research studies usually require quite large samples in order to provide statistical power) and about what constitutes legitimate problems, solutions and criteria of 'proof' (Kuhn 1970).

Both approaches have their strengths and weaknesses (Table 4), and it is important to select the approach which is appropriate for the research question, or a combination of approaches which together provide complementary information.

An increasing recognition of how both paradigms can be applied successfully to medical education scholarship, if done so guided by theory and reflection, is now apparent in the medical education literature (e.g. Ringsted et al. 2011).

Coming back to the main focus of this Guide, note that different research paradigms generate different types of data. Usually, data from qualitative studies using data collection methods such as interviews, focus groups, video diaries and so on cannot be reduced into very simple numerical form for analysis whereas categorical or qualitative data from a quantitative study can be. For example, if a respondent on a survey is asked for their dietary preference, and answers 'vegetarian', this is can legitimately be coded into a number (e.g. where 0 indicates non-vegetarian and 1 indicates vegetarian) and subjected to statistical analysis without losing meaning. Trying to squeeze narratives into boxes may, however, result in the loss of contextualisation and narrative layering. Qualitative researchers think that to do so sublimates the very qualities that make qualitative data distinctive whereas other researchers think presenting qualitative data within a scientific construct facilitates accessibility.

Box 1. Research paradigms.

Research methodology falls into two main camps: quantitative and qualitative methodology. Quantitative approaches are used to test a hypothesis, to answer questions about 'how much' or 'how many'. Qualitative approaches seek answers to the 'what', 'how' or 'why' of a phenomenon, exploring people's attitudes, behaviours, and so on. Quantitative methods tend to generate numerical data, while qualitative methods tend to generate language data (written or oral). In quantitative research, the research question can be exploratory or explanatory, while qualitative research is always exploratory.

There is an ongoing 'paradigm' war (Gage, 1989) in medical education. Clinicians working in medical education have been taught and trained in the scientific model. This is based on quantitative research - gathering empirical and measurable evidence through experiments, hypothesis testing (e.g., 'is this treatment or intervention effective?'), controlling variables other than the one(s) under study so that any statistically significant difference must be a consequence of the treatment/intervention, and producing objective data which can be verified by other scientists reproducing the study. Randomised controlled trials (RCTs) are the gold standard research methodology of this paradigm. If one believes science is the only reliable source of knowledge, then the positivist, or scientific, model is therefore the appropriate model for all research, including educational research. However, social scientists working in medical education have been taught and trained in quite a different model, one that focuses on the human social life and which requires quite a different approach to understand what people do and why. In methodological terms, social scientists argue that we cannot understand why people do what they do, or why particular institutions exist and operate in characteristic ways, without grasping how those involved interpret and make sense of their world: in other words without understanding the distinctive nature of their perceptions, beliefs, attitudes. There is no 'gold standard' methodology; rather social scientists tend to use qualitative methods to explore the phenomenon under question.

Table 4. An overview of some of the strengths and weaknesses of quantitative and qualitative research methods.

Quantitative research methods		Qualitative research methods	
Strengths	Weaknesses	Strengths	Weaknesses
Precision - through quantitative and reliable measurement.	Because of the complexity of human experience it is difficult to rule out or control all the variables.	Qualitative descriptions can play the important role of suggesting possible relationships, causes, effects and dynamic processes.	Contexts, situations, events, conditions and interactions cannot be replicated to any extent nor can generalisations be made to a wider context than the one studied with any confidence.
Control - through sampling and design.	Because of human agency people do not all respond in the same ways as inert matter in the physical science.	Because of close researcher involvement, the researcher gains an insider's view of the field. This allows the researcher to find issues that are often missed (such as subtleties and complexities) by the scientific, more positivistic enquiries.	Researcher's presence has a profound effect on the subjects of study.
Ability to produce causality statements, through the use of controlled experiments	It leads to the assumption that facts are true and the same for all people all of the time.	Qualitative research adds flesh and blood to analysis.	Issues of anonymity and confidentiality present problems when selecting findings.
Statistical techniques allow for sophisticated analyses.	It is not totally objective because the researcher is subjectively involved in the very choice of a problem as worthy of investigation and in the interpretation of the results.		The viewpoints of both researcher and participants have to be identified and elucidated because of issues of bias.

Unsurprisingly, it is the quantitative researchers, those working within a scientific model, who support the latter approach.

There are many books, research papers and AMEE guides available that help determine whether or not your research question is best answered by qualitative or quantitative research paradigms and methods, or if a combination of these approaches (a mixed method study) is appropriate, and how to plan your research accordingly. To find out more about qualitative and quantitative paradigms, their differences, strengths and weaknesses, the following classic texts are helpful: Firestone (1987), Bryman (1984), Norman & Streiner (2000) or Denzin & Lincolm (2001). More recently, Ringsted et al. (2011) provide a useful Guide for choosing a research approach that is appropriate to the purpose of the study while considering the individual researcher's preferences and the contextual possibilities and constraints. Cook (2012) provides a very accessible overview of randomized controlled trials and meta-analyses, and their role in medical education research. Patricio & Vaz Carneiro (2012) compare the evidence produced by systematic reviews of evidence in medical education (BEME Reviews) and clinical medicine (Cochrane Reviews). Other authors demonstrate the use of different qualitative methodologies in medical education research (e.g. Carroll et al., 2008; Cleland et al., 2008; Todres et al., 2012). This list is not meant to be exhaustive but these books and papers introduce some different methodologies for you to consider.

In conclusion, deciding the nature of your research and understanding the nature of your data are crucial when it comes to considering how to manage and analyse your data. Some types of data are appropriate for some types of analysis, others are not. For example, it does not make sense

to compute the average (or mean) of nominal data; imagine computing the mean of gender! After introducing different types of database and what you may wish to consider when deciding which to use, we go on to talk about preparing the data in your database for analysis.

Types of databases

In this section, we introduce different types of databases and discuss their pros and cons for research purposes.

What is the difference between a database and a spreadsheet? Why can't I just use a spreadsheet for my research project? Do I need a statistical or qualitative software package? These are very common questions, which apply equally to quantitative and qualitative research, causing a lot of researchers to scratch their heads. To help, this section weighs up the pros and cons of different types of software: spreadsheet, database, statistical and qualitative software packages.

Spreadsheet software

Many people will already be familiar with spreadsheet packages such as Microsoft Excel™. These feature individual cells defined by a set of rows and columns that can be used to enter data and have useful features such as the ability to quickly take the sum or average of a set of numbers. They are often able to produce graphs very quickly and are also great for rough working and for 'what-if' scenarios – how does changing one cell affect a whole system.

However, spreadsheets are not necessarily useful for a large research project. There are several reasons for this. Firstly, a spreadsheet usually consists of just one table whereas a database (see later) may contain a number of related tables which are integrated. It is also extremely easy to enter data in a format that is then difficult to analyse. For example, if there is a specific reason why a particular numeric value is unavailable, it may be tempting to write a comment in that cell of the spreadsheet. If, on the other hand, rather than a spreadsheet you use a statistical software package (see later), it will require you to keep text and numeric data separate, which means fewer hitches when you come to the analysis stage. Although some spreadsheets offer statistical functions, these are not generally adequate for most research project analyses (spreadsheets are not designed to do complex statistical analysis) and instead data should be imported into a statistical software package. However, you can also use spreadsheets when dealing with qualitative data, especially when using a framework approach where you want to be able to arrange, display and map out the data into a more easily digestible format (e.g. Miles & Huberman 1994).

Database software

Specific database software (such as Microsoft Access™) has many advantages for data entry, data checking, display of data and the ability to store data in complex formats. Some database software incorporates automatic data checks, e.g. they can be set up so that it is impossible to enter values that are out of a sensible range, such as 'alien' when the only two possible values should be 'male' or 'female', or 200 for someone's age.

Database software packages are powerful and can allow complex data that are hierarchical (one to many), for example, when each student has marks in more than one subject and each student has taken a variable number of tests or exams in each subject.

Database software also organises information on a particular subject for retrieval. Data can be retrieved through methods such as asking questions of the data (querying), sorting or filtering, and pulling information into a formatted report that can be printed. Databases also check certain fields, when instructed, to prevent unique identifiers such as patient numbers from being duplicated. This duplication check is not available in spreadsheet software.

Databases are actually much more powerful than spreadsheets in the way users are able to manipulate data. Here are just a few of the actions that can be performed on a database that would be difficult, if not impossible, to perform on a spreadsheet:

- Retrieve all records that match certain criteria
- Update records in bulk
- Cross-reference records in different tables
- Perform complex aggregate calculations

For these reasons, databases may be most useful when handling large amounts of information.

The main disadvantage of using database software for a research project is that it is optimised for data entry and routine data management and not for statistical analysis. For most research projects this means that data will need to be exported to another software package, specifically developed for analysis. However, data may not come through in an optimal format. Often variable and value labels may not be transferred correctly. For example, you might enter your data into Package A, and then try to convert it to Package B and find out that you used the latest version of Package A, but your version of Package B has trouble reading the latest Package A files.

Statistical and qualitative data management and analysis software

Statistical software. If you are working with quantitative (numerical) data and need to perform statistical analysis there is a third option: statistical software. The main advantage of entering data directly into a statistical package is that it removes the need to transfer the data for your analysis. As previously stated, transferring data from one package to another can be difficult. In addition, some data, such as dates, are notoriously difficult to transfer. Setting up data entry in a statistical package also forces you to think about how the data will be analysed from the outset. Often too many unnecessary data items are collected for a research project, or they are obtained in a format that is subsequently difficult to analyse.

In short, if the project involves statistical analysis, enter data directly into a statistical package from the start, don't use spreadsheet or database software. This will force you to think about the types of data you are using (see earlier) and will save time in the long run.

Qualitative software. But what if you are working with qualitative (word or visual image) data? If your data takes the form of 'words' and text, or documents and video clips, there are specific qualitative research software packages available to facilitate data management and analysis.

Qualitative research databases facilitate the interpretation of qualitative data through the coding of themes, concepts, processes, contexts, etc., in order to build explanations, theories or to test or enlarge on a theory. Qualitative software packages are useful for managing large amounts of qualitative data and can save time in terms of manual sorting and organising data. You can usually do the following with a qualitative database: colour your written text or highlight text segments using a marker; use drag and drop coding; get an overview of various object types like primary documents, quotations (i.e. coded segments), codes, memos and saved network views; get a full overview of all codes or memos at any time; manage (sort, rename, merge, delete) codes conveniently; always see your coding; label fine-grained units of analysis (e.g. text characters, image pixels); use colour-coded and grouped codes and so on. Qualitative databases make managing and analysis a large amount of text and/or visual data much, much easier.

However, a word of caution: if you are not familiar with these packages, they can take a while to learn (this is no different from a quantitative database). For smaller qualitative studies (e.g. small pilot study with up to ten interviews) you might wish to consider working with a more manual approach using a spreadsheet to help organise and display your data. A combined approach of Excel™ and Word™, rather than a specific qualitative software package, can work well if your study is relatively small.

Choosing a database

There are a number of considerations when choosing a database (Box 2). Possible users of a database could be researchers, students, educators, administrators, statisticians (if a numerical database) or those entering the data. They could have no experience using databases, some experience of using different software or some experience using databases in the chosen software. For large complex projects it may be advisable to include an experienced database designer as part of the research team.

Box 2. Choosing a database.

The choice of database depends on a number of factors. The main considerations are:

- What is available? Some software packages may require an expensive licence fee.
- Who will be using the database and how experienced are they in using the chosen software? The general ease of use of a database and possible requirement for training is important.
- What will the database be used for?
- What resources (e.g., staff time, money) can be spent developing it?
- Who will enter or check the data? What skills and experience do they have? Will there be more than one person entering data at the same time?

The length of time that this database will be in use, and the importance of the data [that will be contained within the database], must be fully accounted for when deciding how many resources to set aside for buying or creating the database. If a database will only be used for one week to analyse results from a pilot formative test, then a very simple database may suffice. Conversely, a database designed to contain course evaluation data for the next 20 years will need to be carefully designed, with thought given to possible future requirements. Changing the design of a database once it is in use is possible, but not always easy to do and may cause inconsistencies as the data are collected over time.

You might be wondering how software can code and analyse data from, for example, interview or focus group transcriptions – it may be easier to see how statistical software can carry out computations on numerical data. The bottom line is that both statistical and qualitative research software packages are designed brilliantly – to do what you tell them to do (not necessarily what you want them to do)! The manuals and websites of software packages provide much invaluable information to help you in this, but do seek out other people in your institution who have used the package before, and ask them for any handy hints or advice they can provide.

As before, it is best to start as you mean to go on. For a large project, specific database software offers many advantages. If you intend to carry out statistical or qualitative analysis on your data, use the appropriate specialist research software package from the start. Basically, if you are working with numbers, or data which can be sensibly coded into numerical form (see later), you need a database that is designed to store and analyse numerical data. On the other hand, if your study design is qualitative and hence your data takes the form of 'words' and text, or images and visual material, you may want to use a specialist qualitative database to facilitate data management and analysis.

We cannot recommend specific databases, but many are available commercially and are used widely by universities. The use of a commercial database requires a licence, which has a cost attached. The choice of database may be dictated by the resources of your institution, your personal preference and/or what support is available locally.

Working with research databases

Quantitative research data

As discussed previously, in quantitative research, databases hold quantitative or categorical data that has usually been translated into numerical form for the purposes of storage in a database and for statistical analysis. The data may have been collected via a questionnaire survey of, for example, course evaluation, where students may have answered on a 1–5 scale where one is very poor and 5 is very good. A clinical example would be patient responses on a scale of never, once a year, once a month, once a week, every day. In this section, we explain some of the fundamental steps in setting up your database and entering your data so it is ready for analysis.

	Student ID	Year of Study	Gender	Test 1	Test 2	Test 3
1	13794	2	Female	18	16	19
2	18247	3	Male	11	13	14
3	20128	2	Male	15	17	13
4	12873	2	Female	15	14	16
5	19270	3	Male	9	10	6
6	18022	3	Female	11	12	18

Figure 2. Example of records/cases and variables.

Cases and variables

First we introduce two fundamental concepts: records/cases and variables. Usually a case or record is signified by a row in your database, while a variable is a column. We have provided an illustrative example (Figure 2). In this example the rows (the 'cases') represent students, while the columns represent 'variables', i.e. attributes of the students such as their ID number, their gender or their test results.

A 'record' or 'case' can be thought of as a form divided up into areas into which are placed specific types of information. There can be any number of records/cases in a database. Each could typically correspond to a person (as in student records), but equally can correspond to an 'individual case' of a chunk of information such as a response to one interview question or an episode from a transcript. A range of operations can be performed on a set of such records, a process resembling filling in, shuffling and sorting a set of file cards. These operations include sorting cases/records, finding records of a given type, classifying and coding the information in a given field, finding instances of a given type of information, counting instances and so on. In this way, an electronic database can display information in whatever way one chooses, limited mainly by practical considerations.

A variable, on the other hand, is a symbolic name representing some attribute of the case (a piece of data that may vary from person to person, record to record). For example, a variable might be: score on the Year 1 OSCE exam, age, gender, score on a written exam or attendance on the wards.

Sometimes multiple variables are required to measure one aspect – for example, if students receive a series of tests on the same subject, a separate variable is required for each test. You may have a choice between setting up your data using different formats (Figure 3). In long format, each row (or case) represents the result of a particular test, in a particular subject, for a particular student. For example, in the first (top) format presented in Figure 3, in the first row of data, we see that student 1 received a score of 16, for test one, in mathematics. In an intermediate format, each row represents the results of multiple tests, in a particular subject, for a particular student. For example, in the second (bottom) format presented in Figure 3, in the first row of data, we see the scores that student 1 received for tests one to four, in mathematics. There is also wide format (not pictured) where each row represents the result of multiple tests, in multiple subjects, for a particular student. For example, in the row of data, you would see the scores that student 1 received for four tests in mathematics and another four tests in english.

In questionnaire design, a situation occurs that often catches people out. Frequently, there is a mixture of types of questions, some with the instruction 'tick one box' and others 'tick all that apply'. When setting up a database for this type of questionnaire, it is easy to forget that those of the latter type require a separate variable for each possible response option, whereas for the 'tick one box' variety, only one variable is needed. It is important to consider how the database may be set up as you design your instrument (i.e. survey) for these reasons!

Variable names and value labels

As discussed earlier, in most systems the columns of the matrix will represent variables (e.g. student names or the marks in a particular exam) and the rows will represent cases (e.g. individual students). Labelling your variables clearly is critical – someone else might take over the project and need to make sense of your labels or you yourself may have a period of time away from the project, and not understand your own labels when you get back to it (believe us, this happens more often than you might think!). The same considerations apply to value labels - numeric codes for categorical variables, e.g., 0 for male, 1 for female. It is very important to document exactly what the codes stand for, as these can easily be forgotten. See Box 3 for some 'handy hints' for variable and values labelling.

How variables are named is very important. They should be kept relatively short, for ease of use. On the other hand, they also need to be informative. A variable containing the grade for an English course might be called 'GradeEnglish' to help distinguish it from the variable for a mathematics course, 'GradeMaths'. This tells a user of the database more information, and will lead to less confusion, than simply calling the variables 'Grade1' and 'Grade2'. However, note that in the past, some software packages limited variable names to no more than eight characters and even today it is usually not possible to include spaces or special characters as part of a variable name.

To work around these problems, some software packages allow both a variable name and a variable label, which may be longer, allow spaces and special characters and which provides a more exact description of what the variable represents.

Student	Subject	Test No	Score
1	Mathematics	1	16
1	Mathematics	2	11
1	Mathematics	4	12
2	Mathematics	1	13
2	Mathematics	2	12
2	Mathematics	3	13
2	Mathematics	4	9
2	English	1	15
2	English	2	14
3	Mathematics	1	19
3	Mathematics	2	18
3	Mathematics	3	16
3	Mathematics	4	18
3	English	1	17
3	English	2	20
3	English	3	17
3	English	4	16
4	English	1	12
4	English	3	14
4	English	4	15

Student	Subject	Test 1	Test 2	Test 3	Test 4
1	Mathematics	16	11		12
2	Mathematics	13	12	13	9
2	English	15	14		
3	Mathematics	19	18	16	18
3	English	17	20	17	16
4	English	12		14	15

Figure 3. Two ways of displaying the same data: long format (top) or intermediate format (bottom).

Box 3. Handy hints.

Always document your variable and value labels – don't rely on memory or assume it's obvious.
Be consistent with your value labels – keep No and Yes the same (e.g., 0 and 1) wherever these are in the database.
Codes 0 and 1 are often preferred for value labels, rather than 1 and 2, as this makes the interpretation of some types of statistical analysis easier.

This advice may seem trivial and unnecessary. Nevertheless, several different people may use this database, with varying levels of familiarity of its design. What may seem obvious to one person may not be to another. In addition, a database may not be used for a while, before it is required again. Individuals may forget what piece of information is contained within a variable.

It is recommended that variables be named consistently. For example, the variable for question 1, part 1 may be called 'Q1P1'. The variable for question 2, part 2 should then be called 'Q1P2' and not 'P2Q1', say.

To avoid any confusion, ensure the meaning of variables and value codes are made explicit in a formal coding sheet or data dictionary that lists the exact meaning of every variable and code number used (Figure 4). If your software allows this, it may be acceptable to document this within the program itself, but be aware that value and variable labels may be lost when the data are transferred to another package.

Unique identifiers

Often a database is used by many different people, for a variety of reasons. Access to all the data contained within the database may need to be restricted. Even if this is not the case, it is still good practice to keep confidential information, such as students' names and addresses, separate from other data. An effective way to ensure confidentiality is to create two databases: one for confidential data, the other for non-confidential data.

The database containing all the confidential data should contain a unique identifier (ID) for each record. This same unique ID should also be used in the database containing the rest of the information for that record. This will allow for the data kept in both databases to be linked, if necessary. For example, the names and salaries of staff members could be kept in one database, while the number of students and hours each staff members teaches could be kept in another database, and the two databases could be linked using staff ID numbers. Access to the confidential information should be restricted,

Variable description	Variable Name	Codes
Student ID number	id	
Sex	sex	0 Female
		1 Male
Age at start of course	age	
Interest in choosing a career in general practice	gp	1 Very low
		2 Low
		3 Moderate
		4 High
		Very
		5 high
Interest in choosing a career in surgery	surg	1 Very low
		2 Low
		3 Moderate
		4 High
		Very
		5 high

Figure 4. Example of a coding sheet (data dictionary).

Box 4. Handy hint: Never mix text and numeric data.

However tempting, never mix text and numeric data as this will be time-consuming to sort out. Write notes in separate text variables. Do not write 'N/A' or 'Left exam due to sickness' in the middle of a numeric variable. This will cause problems when you attempt to analyse your data (see below).

either using a password protected database or a secure file location.

This ID should be an alphanumeric variable and it should be something that cannot be linked back to the individual (see the section on Ethics). To illustrate, each record could simply be numbered 1, 2 and so on. Or perhaps S1, A1, S2, A2, could be used to identify the first student on the science course, the first student on arts course, the second student on the science course, the second student on the arts course, and so forth (Box 4).

This unique ID will prove invaluable when checking the data contained in the database for odd or inconsistent values. Without it, several pieces of information, such as a student's name, date of birth and so on, may be required to check copies of records for data entry errors. With one, it may be simpler to compare electronic and hard copies of the same information. The use of a unique identifier is also recommended in terms of confidentiality.

In addition, this ID can be used to link data from different sources together. If a student's personal information, courses taken, and exam results are all kept separately, the ID unique to each student can be used to combine all of their details together into one dataset.

Data entry

Data entry is an area which attracts little attention in the research arena. Data can be entered into a database in different ways. In the early days of computing, data were sometimes entered by a manual system of reading punched pieces of cardboard. Later data had to be input using command syntax. Nowadays, some specialist statistical packages require data in a particular format such as a text file with data separated by commas or spaces, but most data can now be entered using a spreadsheet format by typing data into cells of a matrix. 'Raw' data can sometimes be imported directly, such as saving an interview transcript into a qualitative data analysis (QDA) software package. For some research projects optical scanners can be used to read data directly from questionnaires. If carrying out an online survey,

you may be able to set up the survey so that completed questionnaires are entered automatically into a database, depending on the precise combinations of survey package and database software used.

However, depending on the nature of your project and the resources available to you, you may have to manually enter data from paper questionnaires or exam papers into a database, a tedious and time-consuming task. It is often assumed that this 'comes naturally', but our experience is that some personalities are better suited to meticulous data entry than others. So, when planning your project, consider how will the data be entered? If manually, who will enter the data? Would two heads be better than one (one reading out the data, one entering it)? What skills and experience do they have? Who is going to check the data entry?

Manual data entry is a tedious task and errors are commonplace (Box 5). As a result, it is important to check the quality and accuracy of the data entry prior to the data analysis stage (discussed later).

The gold standard method for data validation would be duplicate data entry, where all records are entered twice and then compared in order to identify any discrepancies between the two versions. However, duplicate data entry is time-consuming and costly; so many projects use a check of a certain percentage of the data entry as a quality assurance method. A variety of percentages, e.g. 5% 10%, 20% checks have been used – the proportion to be checked needs to be decided based on the total number of records and the amount of resources that can be allocated. For many projects, there may not be the resources available for these kinds of data checks, but it is still very important to carry out range and consistency checks, as outlined later.

Box 5. 'Rubbish in, rubbish out' (RIRO).

If rubbish (garbage) is entered into the database, this will obviously have a great effect on what comes out. If this happens and data is not double entered or checked, the likelihood is that the whole procedure will have wasted time, effort and money.

Making data entry as simple and as straightforward as possible is the easiest way to prevent bad data being entered. Data checking or validation is also good practice.

Form design

Often the design of a database follows on directly from the design of a form or questionnaire. It is often possible to include code numbers on a questionnaire, which can be used for data entry. Using database software, it may also be possible to set up data entry so that the format of the data entry screen is exactly the same as the format of the questionnaire itself (see, e.g. Figure 5). As discussed earlier, typically, only one line or record of data will be entered for each person (e.g. lecturer) or item (e.g. assignment) of interest.

Note that it is usually better to make things as easy as possible for data entry. If questions need to be reverse scored, this can be done at the analysis stage.

How to designate missing data?

Missing data are, unfortunately, a common occurrence when working with research databases – however hard you try, it is likely that some information will remain unknown and this could be for a variety of reasons.

When setting up a database you need to consider how missing data will be recorded. The most common situation is just to leave missing values blank and not to enter anything into a cell. It is also possible to assign specific codes, typically 9 or 99 to indicate missing values. If doing so, however, you need to be careful to ensure that actual values of 9 or 99 are not possible for this particular variable and to account for the coding at the analysis stage – e.g. taking an average of a set of numbers that includes values of 99 could give you very misleading results.

Checking data

This section introduces some basic descriptive statistics which are useful for describing your data, to ensure that what has been entered into the database is correct. It is very tempting to jump straight in and start to analyse your data, but this is not recommended. First, you need to clean your data and check for possible errors.

Visual inspection of data is a common method of initial quality control, although this should be no substitute for formal data checks. This is often called the 'eyeball technique' or 'eyeball method' of data assessment (Box 6).

- For each question, please circle the number which best applies to you.

		Agree	Neither agree or disagree	Disagree
J1	I can manage the demands of my course	0	1	2
J2	Sometimes I feel like I cannot cope	0	1	2
J3	I enjoy learning about a variety of topics	0	1	2

Note that it is usually better to make things as easy as possible for data entry. If questions need to be reverse scored, this can be done at the analysis stage

Figure 5. Example of a questionnaire format incorporating coding that can be used for data entry.

Box 6. Eyeballing your data.

In a study of student admission, where the variable name represented the offer of a place, students offered a place were coded '1', those not offered a place were coded '0'. Eyeballing the data indicated a couple of other numbers in the column – a 3 and a 9. This was not possible given there were only two coded categories for that variable. Checking against the raw (paper data) indicated that these were data error entries, which were then corrected.

In addition to be detecting visible differences between groups, it may enable you to spot many other things (drift over time, obvious outliers/typos in your data set). It acts as a sense check for your data, but it is no substitute for formal data checks, especially when the size of the dataset is large. There are two further types of checks that can be made to ensure clean and valid data: *range checks* and *consistency checks*.

Range checks ensure data are within a sensible range, for example, if someone's age had been entered as 200 then it would have to be an error. Consistency checks ensure that pairs of variable are consistent. For example, if someone replies that they are a non-smoker in one question but later admit to smoking 20 cigarettes a day, this would also be a mistake of some kind.

Range checks can be performed by obtaining the maximum and minimum values or histograms (for quantitative variables) or frequency tables (for categorical variables). Consistency checks may be performed using scatter plots or using cross-tabulation tables.

Both these types of checks are just common sense but are often forgotten, leading to potentially misleading results. There is also an advantage if these checks are performed automatically as part of the database structure. Packages such as Microsoft Access™ may be set up so that it is impossible to enter an age below zero or above, say, 120 at the data entry stage. This may reduce the number of errors in the data. It may also, however, be time consuming to set up and it may be difficult to know at the outset what ranges are sensible. Returning to the example on age, it would be clearly impossible for anyone to be aged -1 but it is not possible to set a firm upper limit on this variable. Another option within some databases would be to set up queries to highlight those records with out of range values. Getting into a routine of running these daily, or at the end of a batch of entries, means that errors or inconsistencies can be investigated whilst you still have the source data there. Otherwise, if you do find possible errors or inconsistencies, you will need to go back to the source of the data and check this. This may mean checking paper records or going back to the person who supplied you with the data. Such data checks cannot pick up all possible data entry errors, but they should pick up the majority and are recommended before starting your data analysis to ensure high data quality.

Describing data

Graphs are a very clear and effective way of presenting your data. The best way to describe your data will depend on the data type: for categorical data you can display your results using a bar chart. When you have quantitative data you should first graph your data using a histogram: this will show you the overall shape or distribution of your data (Figure 6). For a discussion on the difference between a bar chart and a histogram, see Box 7).

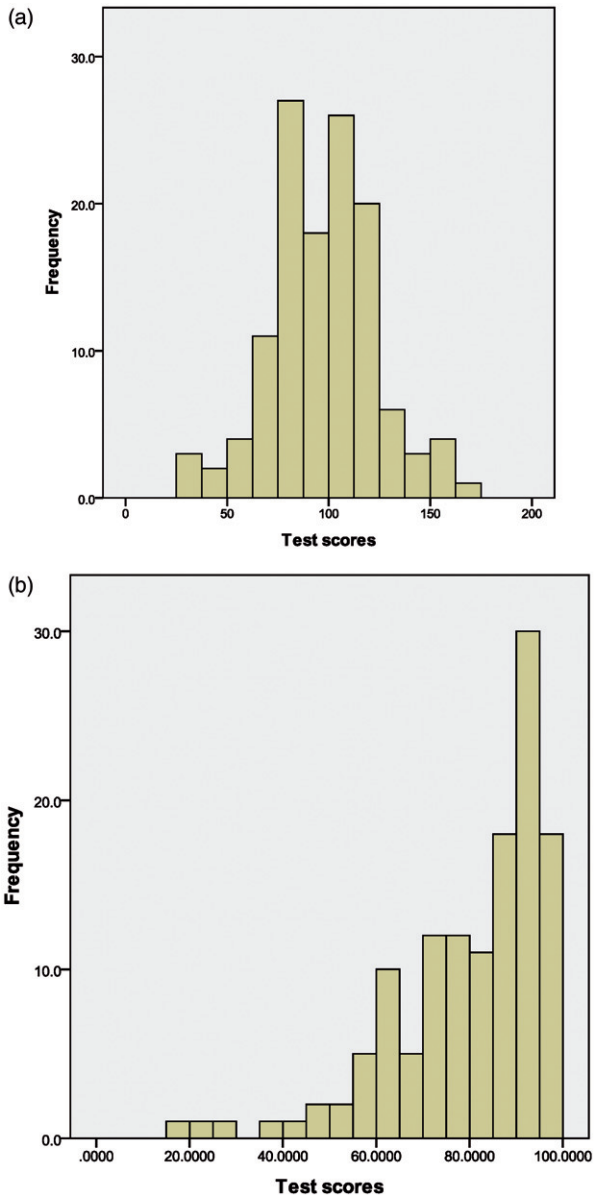


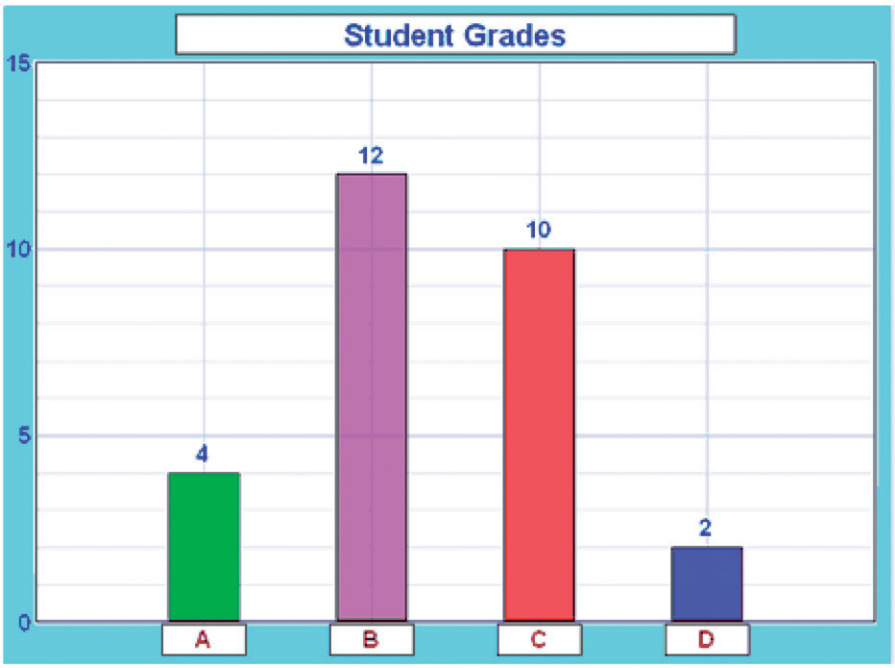
Figure 6. The results of two student tests. (a) The best data descriptors will be mean and standard deviation as the distribution is symmetrical; (b) use median and range as the distribution is skewed.

It is often useful to describe your data using appropriate summary statistics – once again this will depend on the type of data that you have. For quantitative data, however, there is a further consideration – the shape or distribution of the data. What this means is that it is usually best to start by plotting a histogram to help you decide which summaries to present.

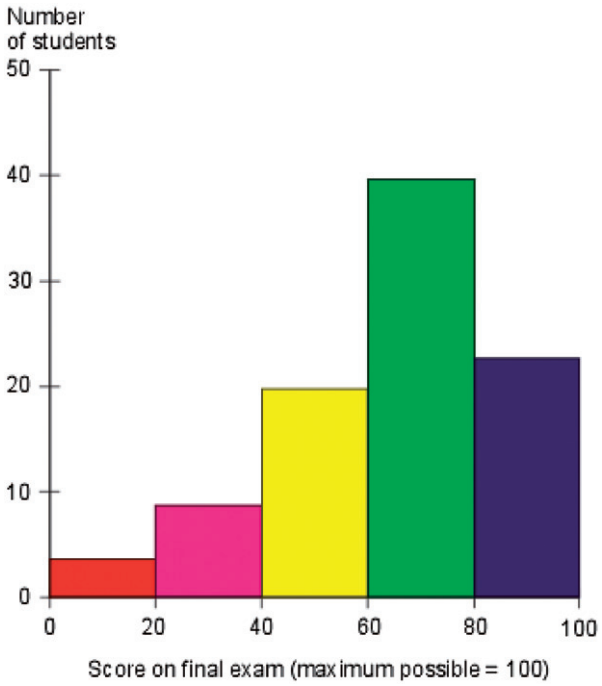
Provided that you have roughly symmetrical data, the mean (the simple average) will be the best single-figure summary of your data. To indicate the spread or dispersion of your data you should use the standard deviation – this is useful

Box 7. The difference between a bar chart and a histogram.

A bar chart is made up of columns plotted on a graph and is used for displaying frequencies of categorical variables where the height of the column indicates the size of the group defined by the column label. An example of a bar chart is provided below. A bar chart is a really good way to show relative sizes: it is easy to see which grades were most and least common, at a glance.



Like a bar chart, a histogram is made up of columns plotted on a graph. Usually, there is no space between adjacent columns. Histograms are used for quantitative variables, where the height of each column indicates the size of the group defined by a range of values. An example of a histogram where each column represents an interval (or 'bin size') of 20 marks is provided below – for presentation purposes the size of this interval may be varied but, provided numbers are reasonably large, the shape of the data distribution should always be similar.



The main difference between bar charts and histograms is that, with bar charts, each column represents a group defined by a categorical variable while, with histograms, each column represents a range of values defined by a quantitative variable.

because most data values will fall within two standard deviations of the mean.

If your data have a *skewed* distribution, the mean and standard deviation will often not be sensible summaries of your data as they will be highly influenced by extreme data values – instead you should use the median (the middle value when your data are arranged in ascending order) as the main summary measure. To describe the spread of your data, you

Box 8. Example of a quantitative research study (Cleland et al., 2008a).

Cohort study on predicting grades: is performance on early MB ChB assessments predictive of later undergraduate grades?

This study aimed to identify whether poor performance in degree assessments early in a UK medical degree course predicts poor performance in later medical degree programme assessments. **(The research question.)**

The study subjects were University of Aberdeen undergraduate MB ChB students graduating in 2003–07 (five cohorts of students). Data on gender, age, funding status (home or overseas), previous qualifications (graduate or not) and marks were routinely collected during the selection and degree assessment processes. **(The data/information.)**

SPSS for Windows Version 15.0 (SPSS Inc., Chicago, IL, USA) was used for data storage and analysis. **(The database.)**

After statistically adjustment for cohort, age, gender, funding source, intercalation and graduate status, [poor] performance (fail and borderline pass) in the Year 2 first semester written examination Principles of Medicine II was found to be a statistically significant predictor of [poor] performance in all subsequent written examinations (all $P < 0.001$). Poor performance in the Year 3 objective structured clinical examination (OSCE) was a significant predictor of poor performance in Year 4 and 5 OSCEs, but there was no statistically significant relationship between essay-based summative assessments and poor performance. Males had statistically significantly poorer performance than females. **(The statistical analysis of the data.)**

We concluded that Examinations taken as early as mid-Year 2 can be used to identify medical students who would benefit from intervention and support. (The conclusion.)

can use the *interquartile range* or the range (the maximum minus the minimum value). In fact, examining the maximum and minimum values is often the quickest way to evaluate if you have any problem values in your dataset.

If you have categorical data you should instead describe your data using frequencies and percentages.

You are now ready to carry out statistical analysis on the quantitative data held in your numerical database! To do so successfully, you need clear research questions, knowledge of statistical analysis and – ideally – the support of a statistician. It is beyond the scope of this Guide to discuss data analysis methods but many useful introductory texts are available (Altman 1991; Bland 2000; Bowling 1997; Norman & Streiner 2000; Petrie & Sabin 2005). We have provided an example of a study using a numerical database, quantitative methods and statistical analysis, where the role of the statistician was central (Box 8).

Qualitative research data

Qualitative research methods tend to generate language or image data. This could be field notes, narratives or other forms of written text such as audio recordings which are then transcribed into written text. Other forms of media such as video recordings, images/photographs and documents (reports, meeting minutes, e-mails) can all be used. The three main types of data in qualitative research are:

(1) Ethnographic field notes – written observations of the field setting that record what is seen, heard or observed.
(2) Interviews designed to elicit stories and accounts, views and attitudes from respondents. Usually interviews are recorded and transcribed, generating large numbers of words. When carefully recorded and transcribed the questions and answers become the data elements.
(3) Focus groups where, usually, three to 10 persons, are asked to address a group of questions for which they have the expertise to provide illuminating information.

The most common forms of qualitative data in medical education research are what people have said during interviews or focus groups. Usually, with appropriate consent from the participants (see later), interviews or focus group interviews are recorded for later transcription and analysis. This can be done using tape recorders with cassettes or digital voice recorders (DVR), which provide better sound quality and recording accuracy, and allow you to download the electrical digital file directly onto a computer, but are often more expensive than tape recorders.

Recording interviews and focus groups is good research practice for a number of reasons. For example:

– If you are taking copious notes, you cannot truly focus in on what is being said in such a way as to manage the interview well.
– Recordings allow you to follow the narrative of an interview, for example, in a group interview who said what, how did people respond to a comment from another participant or from the interviewer. This can be worthy of analysis in itself.
– You can get an impression not just of what participants said (the content of the interview) but also how they said it (in anger, timidly, in a questioning tone of voice and so on).
– Your notes may mean little to you a few weeks later, whereas a recording holds all the information
– A handy hint, no matter which recording device you use, is to have a remote microphone directed at the respondent, and sit near the recording device when asking questions.

Transcribing this data means typing out the text (from interviews, observational notes, memos, etc.) into word processing documents. It is the data from these transcriptions that are managed and later analyzed (most people find it easier to manage data in which has been transcribed rather than working solely from recordings). See Box 9 for an example of transcribed data (from Cleland et al., 2008b).

There is no doubt that recording interviews or focus groups creates a lot of data to manage! If you have collected data from more than a small number of interviews or focus groups, you may want to use a qualitative database (software packages specifically developed for qualitative research) to facilitate this data management and analysis. These are often referred to as Computer-Assisted Qualitative Data Analysis

Box 9. An example of transcribed data (from Cleland et al., 2008b).

Focus group participant 1: *And um predictably the chap was upset, he burst into tears and he was distraught cause this was new feedback for him you know the issue had never been addressed at an earlier stage and that was where the difficulty and the tension came in, uh that it was very hard for us to uh identify for this individual that in our explanation um his performance was substandard in spite of the impression he'd been given over a number of years that he was he was um uh functioning adequately and that maybe a reflection on what uh we've just seen on the slides that poorly performing individuals do not accept feedback readily and it's possible that he had had some sort of feedback that hadn't registered at the requisite level to change his his uh self-appraisal.*

Interviewer: *Yeah. What were the difficulties for you as a medical educator or for your team of medical educators?*

Participant 1: *The main one was just the um that's not the right word really but you know the kind of emotional difficulty that there is in approaching that kind of scenario that you know you're going to you're going to have to disappoint somebody and that's a thing that's uh very uncomfortable so there's a tension in that, you know you've got to do it because that's your responsibility but uh you don't like um hurting people's feelings.*

Participant 2: *And I think that's true at an undergraduate level on the basis that they have already invested huge amounts of time and effort to particularly in the later stages of their undergraduate career and I was going to lead onto just follow on from what xxxxx (participant's name) said in terms of what the problem with the trainee that you alluded to is probably attitudinal rather than intellectual.*

Participant 1: *Absolutely*

Software (CAQDAS; Fielding & Lee 1991). Before we talk more about CAQDAS, it is worth illustrating that many of the issues discussed in the section on numerical or categorical data are also pertinent to managing qualitative research data.

Data entry

As is the case with a numerical database, attention paid to this early stage will greatly help later analysis and interpretation of qualitative research data. Questions you need to consider are – if audio recordings are to be transcribed, who is going to transcribe them? Who is going to check the transcription for quality? What level of detail do you need for your transcriptions, e.g. just spoken word, or do you need to include 'uhms, ahs' or sighs, laughs, etc. For focus group transcriptions, have individual participants been differentiated correctly?

Data checking

If using audio or digital recordings which are then transcribed, whether you carry out the transcription or it is done by a third party, it is extremely important to check your transcripts against the original recordings to check accuracy. If transcription is carried out by a third party, there may be errors in terminology if your topic is quite specialist/specialty orientated. There may also be errors in place and people names (we have had some very funny/strange transcription errors in terms of Scottish town/hospital names). Even if you are transcribing yourself, or someone locally if doing it for you, you must check the transcripts against the recordings as it is very easy to drop a 'not' from a sentence by accident which can then totally change the meaning. The term 'Rubbish In – Rubbish Out' (Box 5) also applies here.

Unique identifiers

Each participant should have a unique identifier (ID). This ID should be an alphanumeric variable as pseudonyms might be considered a form of symbolic violence over the participants (Bourdieu 1991) and it should be something that cannot be linked back to the individual (see earlier and the section on Ethics). Details which could identify the speaker must also be stripped out from the database. For example, in a recent study carried out in Aberdeen, we interviewed physiotherapy, occupational therapy and diagnostic radiography educators about their experiences of assessing students on clinical placement. We realised when looking at the participant background information sheets that all data had to be reported by number (e.g. participant 1, focus group1) only as adding in details such as locality of workplace or gender would have facilitated identification of individuals.

As before, the database containing all the confidential data should be linked to participant background details using only a unique ID and access to the qualitative research database should be restricted.

Data description

Qualitative data analysis (QDA) is the range of processes and procedures whereby we move from the qualitative data that have been collected into some form of explanation, understanding or interpretation of the people and situations under investigation. The idea is to examine the meaningful and symbolic content of qualitative data.

Data describing and analysis in qualitative research tend to overlap. While the purpose of this Guide is not to introduce qualitative research data analysis in depth, it is important to give an overview of the first steps, those akin to describing your data when working with numerical data.

The process of QDA usually involves two things; coding (the identification of themes) and writing. We describe coding here as it most closely resembles data description – writing tends to be more analytic and interpretative and hence beyond the scope of this guide.

Coding. On any given topic and in answers to interviewers' questions, respondent stories may be highly differentiated and varied in content. The researcher must read through much textual data trying to locate and understand different themes or uniformity that emerge from respondent stories. This includes identifying textual material (appropriate quotes) that exemplify the themes (or categories) relevant for the written report. The analyst needs to define categories for the different themes and may assign some code or value to each category. If and when a coding scheme has been generated, the researcher(s) can encode each respondent's answer according

to those themes. In other words, coding is the identification of passages of text (or other meaningful phenomena, such as parts of images) and applying labels to them that indicate they are examples of some thematic idea.

At its simplest, this labelling or coding process enables researchers quickly to retrieve and collect together all the text and other data that they have associated with some thematic idea so that they can be examined together and different cases can be compared in that respect. Coding the data makes it easier to search the data, to make comparisons and to identify any patterns that require further investigation.

Codes can be based on:

- Themes, Topics
- Ideas, Concepts
- Terms, Phrases
- Keywords

found in the data. All passages and chunks that are coded the same way – that is given the same label – have been judged (by the researcher) to be about the same topic, theme, concept, etc. The codes are given meaningful names that give an indication of the idea or concept that underpins the theme or category. Any parts of the data that relate to a code topic are coded with the appropriate label. If a theme is identified from the data that does not quite fit the codes already existing then a new code is created.

As you read through your data set the number of codes will evolve and grow as more topics or themes become apparent. The list of codes thus will help to identify the issues contained in the data set.

During coding, you must keep a master list (i.e. a list of all the descriptors or codes that are developed and used in the research study). Then, the codes are reapplied to new segments of data each time an appropriate segment is encountered. It is surprising how the meanings of codes/themes can evolve during analysis. Therefore, keeping a current description of the code/theme will improve consistency over the whole dataset.

Thus, whereas in a quantitative database you 'code' information in terms of values and variable names, coding in qualitative research means something a little different but similar in that coding is a way of pulling out information for analysis. CAQDAS packages involve tools which can mark and retrieve data through coding the text, such as interview transcripts, field notes, transcribed recordings, documents. Coding involves marking the text in order to tag particular chunks or segments of that text. Codes are thus attached to discrete stretches of data. (How you code the data is up to you and may be influenced by your conceptual or theoretical framework.)

In Figure 7 we present a snapshot of a page of codings from one of our exploratory studies, which looked at identifying communication skills training needs in primary care (Moffat et al. 2007).

Qualitative database software facilitates the attachment of these codes to data; it also allows the researcher to retrieve all instances in the data that share a code. The underlying logic of coding and searching for coded segments differs little, if at all, from that of manual techniques to do the same thing. There is no great conceptual advance over the indexing of typed or even manuscript notes and transcripts, or of marking them physically with code-words, coloured inks and the like. However, using a computer database enables fast and comprehensive searches that can use more than one code-word simultaneously, to facilitate complex searches. The co-occurrence of codings can be an important issue; finding them can be a useful tool. Since the software can handle very large numbers of codings and code words, in purely mechanical terms the computer can help with more comprehensive and more complex code-and-retrieve tasks than can be achieved by manual techniques. Many of the packages also allow the researcher to add notes to the text, which facilitates analysis.

Thematic or conceptual coding is one way of categorising data, but some packages provide alternative means of working with qualitative data. One example of this is the use of linking tools to track non-linear associations, which can be particularly useful for narrative-based approaches (see Silver & Fielding 2008 and Silver & Patashnick 2011 for more discussion on this). CAQDAS packages also provide tools that allow textual data to be explored according to content, i.e. to consider the context within which keywords or phrases are used, which can be useful for approaches interested in the use of language.

The fact that qualitative databases allow for coding alongside the data is hugely helpful in terms of data management, particularly if you have a lot of data to manage. We provide an example of a qualitative research project which used a database to aid data analysis in Box 10.

We cannot provide a full description to qualitative data management and analysis in this Guide, so we direct you to the many textbooks and online resources available to help with this aspect of qualitative data management (e.g. Ryan & Bernard (2003) or Strauss & Corbin (1998)).

Ethics and confidentiality

Any medical education research study involving data collection must follow ethical procedures. These differ by country. For example, the Netherlands does not have rule-based ethics review for education projects. In the UK, however, most medical schools and universities have internal ethics review committees for research involving students or staff – these are different from those committees which review research studies involving patients. In countries without separate review boards for educational projects, the mainstream review boards and ethics committees frequently find it difficult to handle requests for review from the medical education community. On the other hand, medical education journals (the journals you are likely to want to publish your study in!) each have a philosophy-based approach to ethical conduct which requires that authors show how they addressed the spirit of protecting subjects and how they articulated any issues of risk. The journals do not define or police ethical standards but, rather, are clear on what statement is required from the authors about ethical requirements in their country of origin. In other words, they encourage authors to be transparent about issues of research ethics (Brice et al. 2009).

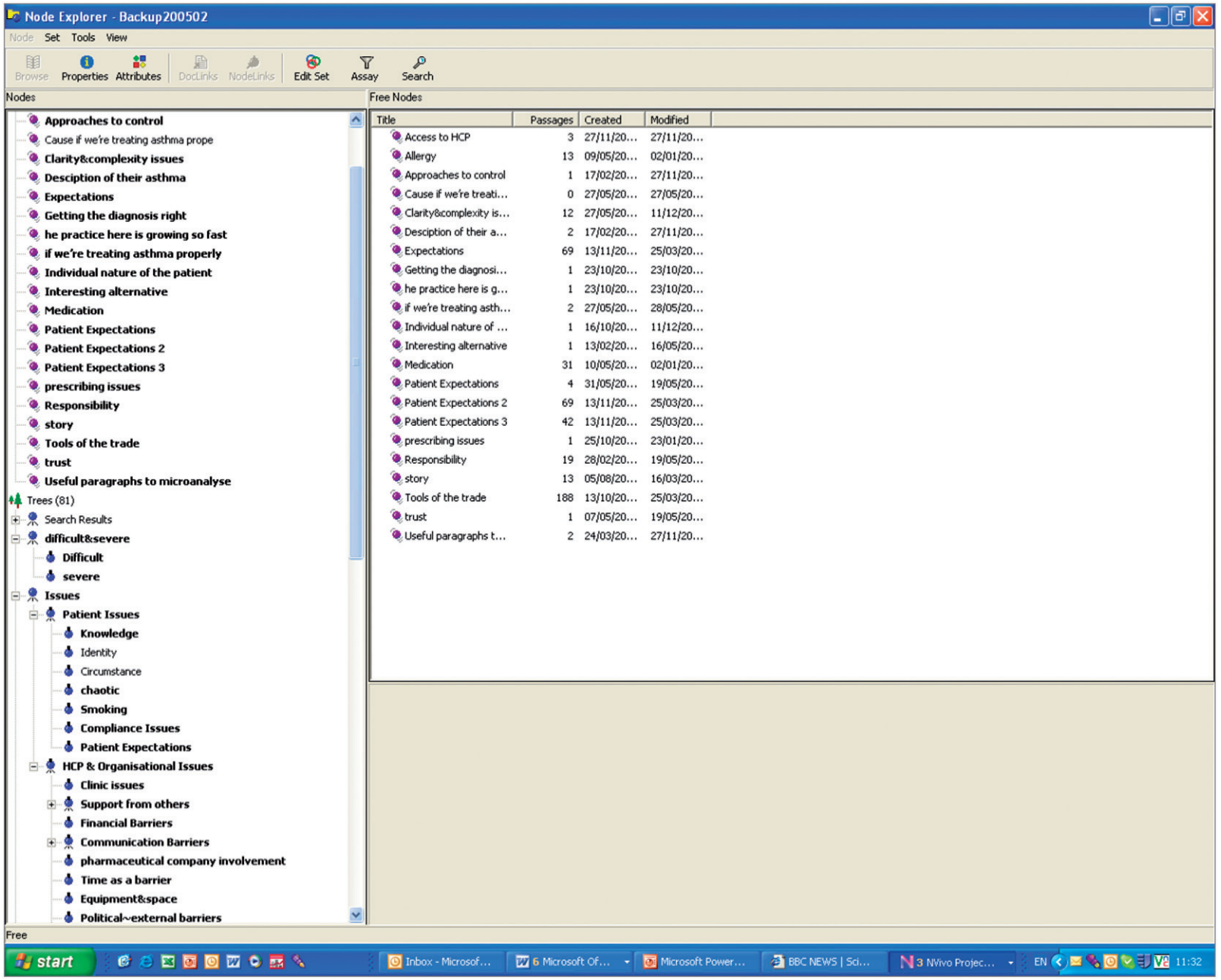


Figure 7. An example of a coding page (Moffat et al. 2007).

Box 10. Example of a qualitative research study (Cleland et al., 2008b).

Is it me or is it them? Factors that influence the passing of underperforming students

This study aimed to explore if 'failure to fail' is an issue for medical educators in the UK, and, if so, what are its determinants? **(The research question)**

The study subjects were medical educators (general practitioners, hospital doctors and non-clinical tutors) from two different UK medical schools who took part in focus group discussions which were recorded and transcribed for analysis **(The data/information).**

The qualitative data analysis software Atlas-ti ^{TM} (Atlas-ti Scientific Software Development GmbH, Berlin, Germany) (http://www.atlasti.com/index.html) was used for data storage and analysis. **(The database).**

Using both theory and data-driven framework analysis, we identified six main themes relevant to the integrative model of behavioural prediction. These were: tutor attitudes towards an individual student; tutor attitudes towards failing a student; normative beliefs and motivation to comply; efficacy beliefs (self-efficacy); skills and knowledge; and environmental constraints **(The data analysis).**

We concluded that many different factors impact on medical educators' failure to report underperformance in students. There are conflicts between these factors and the need to report competence accurately **(The conclusion).**

Our advice is to follow good research practice no matter what the procedures of your country or institution. Broadly speaking, this means:

- Taking necessary steps to ensure that all participants in the research understand the process in which they are to be engaged, including why their participation is necessary, how it will be used and how and to whom it will be reported. This is usually achieved by using a 'participant information sheet' (PIS) outlining the nature of the project and what is involved in taking part. It should be made clear on the PIS that a participant can withdraw from the research for any or no reason, and at any time, without recourse.
- Ensuring potential participants understand and agree to their participation without any duress (i.e. taking part in any research project must be voluntary). Agreeing to participate and knowing what one is agreeing to, can be implied by, for example, questionnaire completion, but often participants are required to sign a separate 'consent form'.
- If you have the funds to use incentives to encourage participation, these should be commensurate with the effort involved in taking part (e.g. providing a sandwich lunch and travel expenses for people taking part in a focus group is reasonable, paying them more than a small sum of money is not). Remember that if you are using incentives, this can influence the research in terms of a bias in sampling or in participant responses.
- The secure, confidential and anonymous treatment of participants' data is considered the norm for the conduct of research. To aid in this, it is desirable to ask the researchers to sign a code of conduct or equivalent, to agree to maintain confidentiality and security. The database should be password protected or held on a [desk-top or] computer which remains within the institution – laptops are often stolen! Any database or data file which has student/doctor identifiable details on it should not be transferred between computers on a memory stick or disk, unless this is encrypted. It is good practice to keep confidential information, such as students' names and addresses, separate from other data.
- You must never send identifiable data outside your institution electronically (unless encrypted) or in paper format. If you are doing a two-centre or multi-centre study where the database has to be sent elsewhere, or you are receiving data from another institution to merge with a database you are holding, the project must be organised so that shared data does not contain student/doctor identifying details. This is good research practice and pertains to any data sharing, not just that held in a database.
- Once the project is finished, the database should be locked and stored (archived) as per institutional guide lines.
- Similarly, researchers must ensure that the form of any publication does not directly or indirectly lead to a breach of agreed confidentiality and anonymity.
- And, particularly relevant to education research, researchers must comply with the legal requirements of their county in relation to the storage and use of personal data. People are entitled to know how and why their personal data is being stored, to what uses it is being put and to whom it may be made available (this information should be included in your PIS).

As discussed earlier, often a database is used by many different people, for a variety of reasons. Access to all the data contained within the database may need to be restricted. Even if this is not the case, access to the confidential information should be restricted. This is usually via password protected files and computers. In addition, the database containing these data should be kept in a secure location.

Using routinely collected data for research purposes

Much data are collected about medical students and doctors that does not relate to specific research projects planned in advance. This concerns data that are routinely collected by medical schools for other purposes (Table 5).

You may want to apply a research question to a database held by the medical school for administrative purposes. If analysis of a routine database will address your research question, as a very rough guide, and bearing in mind the data protection legislation and principles for your country, you are likely to be able to access routine medical education data legitimately if you are a contracted Faculty member (but check the specific rules and regulations of your institution!). If you are using a researcher for data checking and analysis, it is desirable to ask the researcher to sign a code of conduct or equivalent. Alternatively, you could ensure that all person

Table 5. Common uses for databases in undergraduate medical education.

Student records – personal data, home address, age, next-of-kin, etc
Admissions data
Attendance at teaching events
Submission of work
Examination results
Course evaluation/feedback data
Progression (including time out of programme, resits)
etc.

identifying details are removed from the routine database before it is used for research purposes.

This information is invariably collected on a non-consented basis and, as such, there may be ethical issues in using it for research purposes. In other words, an existing database will not have been set up originally for research purposes but will hold routine data – admissions or assessment data for example. In this case, your use of the database is secondary to its main purpose.

There have been long-standing debates as to whether or not use of routine databases in medical education constitutes research or evaluation (see McLachlan and McHarg 2005, for an overview of this debate). Top of page Abstract

Morrison and Prideaux (2001) proposed that research is 'aimed at producing generalisable results to be published in the refereed literature'. Research is intended to benefit a general, non-specified audience, while evaluation is addressed to a particular (and specified) constituency or constituencies. In contrast, evaluation tends to be for local purposes – is our teaching acceptable to students, how does one examination compare to another in terms of difficulty, etc.? However, data gathered for evaluation or other purposes may subsequently be appreciated as generalisable (e.g. Cleland et al. 2008a). The lack of necessity for formal ethical permission from a local research ethics board (REB), including informed consent, for evaluation means that data that are subsequently reclassified as of research interest 'lie in ethical limbo' (McLachlan & McHarg 2005). (Where you are creating a database purely for research purposes, as part of a planned study, the situation is less disputable: you must follow the ethics procedures of your country/institution for educational research.)

While some medical educators feel that subjecting education research to the level of scrutiny of a REB is akin to 'using a sledgehammer to crack a nut' due to the comparatively low risk involved for participants in education studies (Pugsley & Dornan 2007), our view is that it is unethical not to do so. Our own experiences are very positive and often a simple, pre-application query is enough for the Chairman of a REB to give a view as to whether or not a full application is required. Many medical school and universities now have an internal ethics Board or Committee, established to deal with 'low risk' studies and populations, where the procedures are more straightforward (reflecting the low risk of this type of study and the fact that the populations under study tend not to be vulnerable). Furthermore, to publish your work, you may be required to provide evidence that you sought ethical permission (Brice et al. 2009).

We have tried to give some broad guidance on the ethical issues involved in educational research and use of routine databases in educational research. Further reading on this topic includes the following. A useful guide for seeking ethical approval for education research is 'Twelve tips for ethical approval for research in health professions education' by Egan-Lee et al. (2011). A recent editorial in the journal *Medical Education* (Eikelboom et al. 2012) presents a framework for the ethics review of education research and see also Kanter (2009) for the journal *Academic Medicine*'s policy on studies involving human participants. Ten Cate (2009) provides a useful editorial on why the ethics of medical education differ from those of medical research (see also Eva 2009) and the American Educational Research Association (AERA) provides a Code of Ethics for educational research (2011).

Conclusion

This AMEE Guide is intended as an introduction to research using databases in medical education. In addition to outlining many basic principles of research using databases, from setting up to describing your data, it presents an overview of the variety of research data and methodological approaches which are suitable for database research. We have tried to explain the various steps in planning and setting up a well-designed research database and how ethical and careful planning can greatly assist in achieving and providing accurate data management, ready for analysis, whether you are using qualitative or quantitative database software, whichever is appropriate for your research question. We have supported the content with a combination of 'seminal' and more recent references. We have aimed for a Guide which is useful to all researchers, not just those with (relatively) plentiful resources. We hope readers have gained insight into how different types of data can be used in furthering the goal of extending the knowledge and understanding of medical education. Future AMEE Guides will address the next step of data analysis in considerably greater detail.

Acknowledgements

Our thanks to Dr Lorna Aucott and Katie Wilde, University of Aberdeen, Professor Trevor Gibbs and the reviewers for their helpful comments on the manuscript. We also thank Professor Trevor Gibbs for his support in bringing this Guide to fruition.

***Declaration of interest*:** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Notes on Contributors

Professor JENNIFER CLELAND, BSc, MSc, PhD, D Clin Psychol, is lead for medical education research at the University of Aberdeen, UK.

Dr NEIL SCOTT, MA, MSc, PhD, is a Medical Statistician in the Medical Statistics Team, Institute of Applied Health Sciences, University of Aberdeen, UK.

KIRSTEN HARRILD, BSc, MSc, is a Medical Statistician in the Medical Statistics Team, Institute of Applied Health Sciences, University of Aberdeen, UK.

Dr MANDY MOFFAT, BSc, PhD, is a Lecturer in medical education in the Division of Medical and Dental Education, University of Aberdeen, UK.

References

Altman DG. 1991. Practical statistics for medical research. London: Chapman & Hall.

American Educational Research Association. 2011. Code of ethics, american educational research association and Australian council for educational research. Educ Res 40:145–156.

Bland M. 2000. An introduction to medical statistics. Oxford: Oxford University Press.

Bourdieu P. 1991. Language and symbolic power. Boston: Harvard University Press.

Bowling A. 1997. Research methods in health. Buckingham: Open University Press.

Brice J, Bligh J, Bordage G, Colliver J, Cook D, Eva KW, Harden R, Kanter SL, Norman GR. 2009. Publishing ethics in medical education journals. Acad Med 84:S132–S134.

Bryman A. 1984. The debate about quantitative and qualitative research: A question of method or epistemology? Brit J Sociol 35:75–92.

Carroll K, Iedema R, Kerridge R. 2008. Reshaping ICU ward round practices using video-reflexive ethnography. Qual Health Res 18:380–390.

Cleland JA, Milne A, Sinclair HK, Lee AJ. 2008a. Predicting performance cohort study: Is performance on early MBChB assessments predictive of later undergraduate grades? Med Educ 42:676–683.

Cleland JA, Knight L, Rees C, Tracey S, Bond CB. 2008b. 'Is it me or is it them?' Factors influencing assessors' failure to report underperformance in medical students. Med Educ 42:800–809.

Cook D. 2012. Randomized controlled trials and meta-analysis in medical education: What role do they play? Med Teach 34:468–447.

Denzin NR, Lincolm YS. 2001. The SAGE handbook of qualitative research. 4th ed. Thousand Oaks, CA: Sage.

Egan-Lee E, Frietag S, LeBlanc V, Baker L, Reeves S. 2011. Twelve tips for ethical approval for research in health professions education. Med Teach 33:268–272.

Eikelboom JI, ten Cate OTJ, Jaarsma D, Raat JAN, Schuwirth L, van Delden J.M. 2012. A framework for the ethics review of education research. Med Educ 46:728–737.

Eva K. 2009. Research ethics requirements for medical education. Med Educ 43:194–195.

Fielding N, Lee R. 1991. Using computers in qualitative research. London: Sage.

Firestone WA. 1987. Meaning in method: The rhetoric of quantitative and qualitative research. Educ Res 16:16–21.

Gage NL. 1989. The paradigm wars and their aftermath. Educ Res 18:4–10.

Guba EG. 1978. Towards a methodology of naturalistic inquiry in educational evaluation. Los Angeles, CA: Centre for the Study of Evaluation.

Kanter S. 2009. Ethical approval for studies involving human participants: Academic Medicine's new policy. Acad Med 84:149–150.

Kuhn TS. 1970. The structure of scientific revolutions. 2nd ed. Chicago: University of Chicago Press.

McLachlan JC, McHarg J. 2005. Ethical permission for the publication of routinely collected data. Med Educ 39:944–948.

Miles MB, Huberman AM. 1994. Qualitative data analysis: An expanded sourcebook. London: Sage.

Moffat M, Cleland J, van der Molen T, Price D. 2007. Poor communication may impair optimal asthma care: A qualitative study. Fam Pract 24:65–70..

Morrison J, Prideaux D. 2001. Ethics approval for research in medical education. Med Educ 35:1008.

Norman GR, Streiner DL. 2000. Biostatistics – The bare essentials. 2nd ed. Hamilton: BC Decker.

Patricio M, Vaz Carneiro A. 2012. Systematic reviews of evidence in medical education (BEME Reviews) and clinical medicine (Cochrane Reviews): Is the nature of evidence similar? Med Teach 34:474–482.

Petrie A, Sabin C. 2005. Medical statistics at a glance. 2nd ed. Malden, MA: Blackwell Publishing.

Pugsley L, Dornan T. 2007. Using a sledgehammer to crack a nut: Clinical ethics review and medical education research projects. Med Educ 41:726–728.

Ringsted C, Hodges B, Scherpier A. 2011. AMEE Guide 56 research in medical education. 'The research compass': An introduction to research in medical education. AMEE. Med Teach 33:695–709.

Ryan GW,, Bernard HR. 2003. 'Techniques to identify themes. Field Methods 15(1):85–109.

Silver C, Fielding N. 2008. Using computer packages in qualitative research. In: Willig C, Stainton-Rogers W, editors. The Sage handbook of qualitative research in psychology. London: Sage Publications. pp 334–351.

Silver C, Patashnick P. 2011. 'Finding fidelity: Advancing audiovisual analysis using software'. The KWALON Experiment: Discussions on Qualitative Data Analysis Software by Developers and Users, FQS, Vol 12, No 1. pp. 334–351.

Strauss A, Corbin J. 1998. Basics of qualitative research. Grounded theory procedures and techniques. 2nd ed. Newbury Park, CA: Sage.

ten Cate O. 2009. Why the ethics of medical education differ from those of medical research. Med Educ 43:608–610.

Todres M, Tsimtsiou Z, Sidhu K, Stephenson A, Jones R. 2012. Medical students' perceptions of the factors influencing their academic performance: An exploratory interview study with high-achieving and re-sitting medical students. Med Teach 34:e325–e331.