

An Introduction to Research for Primary Dental Care Clinicians

Part 7: Stage 8. Collecting Data

Trevor M Johnson

Introduction

This paper, the seventh in the series, will address the eighth of the ten stages of a research project suggested in the first paper. The ten suggested stages are:

1. The initial idea (asking a research question).
2. Searching the literature.
3. Refining the research question.
4. Planning the study.
5. Writing a protocol.
6. Obtaining ethical approval and funding.
7. Piloting the methodology and project management.
8. **Collecting data.**
9. Analysing the data.
10. Writing up and disseminating the results.

The previous paper outlined how to pilot the methodology and manage a research project. Once the pilot has taken place and any issues have been identified, then data collection can take place. This paper outlines the principles of data collection. It updates two previous FGDP(UK) research advice sheets: *Data Collection* and *Sampling*.

Stage 8. Collecting Data

This paper is divided into the following sections:

- A. Introduction
- B. Data collection
- C. Data management

D. Sampling, randomisation and validity

E. Final thoughts

F. Suggested further reading

A. Introduction

High-quality data collection and robust data management are crucial in research, irrespective of the study design used. The collection of clean and reliable data enables data analysis to take place. Data collection and management should be carefully considered at the protocol stage. Clean data are character and numerical data that contain no invalid characters or numbers; for example, typographical mistakes or numbers outside the specified range of values. In practice-based research,¹ it is important to pilot data-collection procedures and the instruments used to collect the data before the main study starts, as poor data collection or management will jeopardise the whole project.

Through data, it is possible to look at what lies underneath the effect that is being observed and to use the data to lead the researchers to the underlying truth. A collection of each data value is usually tabulated, in spreadsheet form or in a statistical software package such as the Statistical Package for the Social Sciences (SPSS Inc, Chicago, USA), with each column representing a variable. This collection is the data set. Data may also be described as primary (original) or secondary data, which are derived from primary data.

Data collection is about systematically acquiring, validating, and storing variables that interest the researcher so that the research question may be answered, for example one derived using the PICO format as discussed in the second paper in this series.²

B. Data collection

B1. Data collection in practice-based research

It is recognised that research carried out in primary care dental settings is very likely to be of a different type to that in universities, for example practices are structured primarily to treat patients and research should interfere as little as possible with this. However, academics may still be involved in research in primary care settings, as part of the research team. Universities usually have dedicated methodological and statistical support available that may contribute from the planning stages through to publication of the research.

Evidence from existing dental practice networks indicates that three broad categories of data are collected.³ They are:

1. Data from practitioner surveys that do not involve information from specific patient encounters, such as attitudes to current treatment modalities
2. Data from specific patient encounters, such as assessments of gingival health.
3. Data directly from patients, such as questionnaires.

Data collection is different in these three categories and must be carefully considered at the protocol stage of the research project. It is well worth sending the research protocol for review, if not formal refereeing, so that any errors or omissions are identified before data collection begins, as they may not be rectifiable later.

B2. Collection of quantitative and qualitative data

Research methodology can be separated into two main categories, quantitative research and qualitative research. Quantitative research collects numerical data and is used in clinical and observational studies; for example, collecting data about gingival inflammation, carious surfaces, and pain scores. Standardised tools should be used to collect data (for example the International Caries Diagnosis and Assessment System [ICDAS]⁴ for caries assessment; variables are known and the data are analysed statistically) although this is not always the case. Deductive reasoning is used to reach conclusions about data analysis using appropriate statistical tools. Deductive logic begins with certain premises that allow research hypotheses to be developed and tested. Qualitative research is text- or visually-based and may arise from interviews. Furthermore, in qualitative analysis, the numbers of variables are unknown, data collected are not structured, and analysis depends on searching for themes and inductive reasoning. Although it may be assumed that quantitative research is more widely used in healthcare research, the qualitative approach is an important one and may be used, for example, for impact evaluation of strategies. Qualitative and quantitative research may also work together by using a qualitative approach to explore the meaning of something and then describe it, followed up by quantitative research to formulate a research question that can be tested numerically. Some of the differences between quantitative and qualitative approaches to research are at Table 1.

Table 1: Differences between quantitative and qualitative approaches to research		
	Quantitative approach	Qualitative approach
Research question	Test theory Validate	Build theory Describe
Research process	Tightly structured Known variables Set methodology	Loosely structured Unknown variables Variable methodology
Data collection	Numerical Sample size calculated, often large Specific measurement instruments	Text, image Sample size not calculated, often small Non-standardised interviews
Data analysis	Statistical Deductive reasoning	Looking for themes and patterns Inductive reasoning
Outcome presentation	Statistical Numerical Formal	Words Narrative

B3. Data variables and scales

A variable is when there are two or more values for a characteristic that is under investigation. Variables may be classified into different categories. A continuous variable has an infinite number of possible values. For example, in measuring patients' ages, it would be possible, although unlikely, to establish them down to the nearest second. In contrast, discrete variables have a set number of values; for example, the Löe and Silness Gingival Index (1963)⁵ has four discrete values—0, 1, 2, and 3—for the assessment of gingival inflammation.

Variables in research may also be classified as administrative or research. Administrative variables contain metadata, which describe the content and context of the data and enhance the value of the data stored; for example, they may record dates, times, and personnel involved. Research data are the variables and outcomes from the research activity and comprise the data used for primary analysis.

Data are described in four different scales (nominal, ordinal, interval and ratio; Table 2). When deciding upon data to collect, it is important also to decide to which scale the data belong, as this will also determine the statistical analysis. Generally, data cannot be converted from one scale to another, although ordinal data are often dichotomised. 'Dichotomous' data are data from outcomes that can be divided into two categories (for example, teeth or no teeth), where each participant must be in one or the other category, and cannot be in both.

Table 2: Data scales		
Data scale	Description	Statistical test
Nominal data (categorical data)	Numbers used only to identify different categories and have no meaning in themselves	Chi-square Mode
Ordinal data	Numbers used to show a sequence or order, but do not indicate numerical differences	Median Percentile
Interval data	Numbers reflect standard and equal units of measurement and the size of differences	Mean Standard deviation Regression Analysis of variance
Ratio data	Similar to interval data but with a true zero point	Same as interval data, plus Geometric mean/Harmonic mean Coefficient of variation

C. Data management

C1. Data management and planning

Data management planning is important from the outset and a data manager, who then takes responsibility for the safe management of the data collected, should be appointed at the start of the research project. In a practice-based setting, it will often be the lead clinician who takes

on this role and who has to take decisions about the data to be collected, together with input from the other researchers.

First, what data are needed? This depends on the research method used, which may produce clinical, observational data, questionnaire-based data, historical data collected from existing records. The location of the data, although apparently obvious, should be carefully considered. For example, when conducting a survey of clinical records, obtaining the data may mean gaining ethical approval⁶ and the mechanisms for obtaining data should be fully worked out before proceeding further. Finally, the question of interpreting the data remains, so that the most appropriate tools are used, for example, statistical tests. Research in healthcare often involves three elements: clinical, methodological, and statistical. No single person normally possesses skills in all these elements sufficiently to be able to research effectively without help from at least someone from another discipline.

Data may be collected on paper or electronic forms, although frequently a hybrid method, combining the two, is used. An advantage of electronic data capture is that spurious data are detected at the point of entry and not allowed to corrupt the dataset. The system needs to be designed by the data manager, although an existing and proven system may be used. Data integrity and security need to be addressed. The security of the system is important because confidential information may be stored and access should only be given, as required, to key members of the team. A robust, validated, and securely backed-up electronic database, with off-site secondary storage, should be used, so that key data are protected from unauthorised access and potential loss.

The system needs to be tested at the pilot stage of the research project and not when it goes live for the start of the main study. Piloting the database or databases used is critical; for example, in order to check that it will export to the statistical package used for data analysis and that data are not corrupted.

The data manager should also define the study variables at the outset, with permitted range of values for entry so that data entry is harmonised between the study team members. Codification of data for entry into paper or electronic forms, setting up a data validation process, and demonstrating a data trail, which may be audited are also roles for the data manager.

C2. Data-collection forms

Data-collection forms need careful design so that they are consistent and unambiguous, whether they are paper or electronic. Paper collection forms may be easier to use for small-scale projects although they are more difficult to back-up reliably and data cannot be directly validated on entry in the same way as an electronic form may operate.

Double data entry, also known as double key data entry, is about re-entering data for a second time and is a standard method of quality assurance. Software may be used to validate that the data have been correctly entered and electronic validation will prevent erroneous data being collected. Paper forms should be reviewed for incorrect entries and paper data entries transferred to the electronic database at regular intervals.

Electronic data capture is more widespread and offers advantages for improving data quality, by checking data on entry and allowing web-based, online data entry. A new term, intelligent

data capture, is now in existence and rather than simply automating manual entry systems, technology may be used to design a better system from the outset.⁷

Data cleansing is the removal from the system of data that are incomplete, inaccurate, or irrelevant, thereby removing inconsistencies from the database and returning a set of data that is clean. It differs from data validation, which is usually referred to as the removal of data at the point of entry.

C3. Data assurance and quality

Quality assurance in data collection⁸ is important in any setting, but more so in a practice setting, which does not have all the support that an academic institution may have. As described in the previous paper in this series,¹ training and monitoring of personnel in the research project should be regularly reviewed and changes made as necessary.

The promotion of data quality is a key component of the research process. In electronic systems, as has been seen, validation can occur at the point of data entry, at the so-called 'front end'. However, in a paper system, data are checked at the 'back end' and any alterations made at that point. The electronic system, with its associated metadata, has a defined audit trail already in place that may be missing in the paper system. Web-based data-collection systems can be seen to have advantages⁹ because direct entry and validation are usually inbuilt, so that, for instance, out-of-range values are rejected.

C4. Datasets

A dataset is a collection of data in quantitative research that is normally presented in tabular form, where each value is a datum, a row corresponds to the data values for that record, and the columns express variables. The dataset is the basis for analysis and data may be entered into a spreadsheet or statistical software package.

The National Health Service (NHS) Information Centre (www.ic.nh.uk/statistics-anddata-collections) has many datasets that are available for download and their uses are defined as "the data required to provide information leading to knowledge for care".

C5. Data control and archiving

Data control should include means of managing confidential data, which may be personal information, clinical notes and images, still or video. For ethical and security reasons, key personal information (such as name, date of birth, address, telephone numbers, National Insurance number, NHS number) should be stored separate from the data collected as part of the study. In order for this to happen, a unique identifier should be given to each participant. A unique identifier then allows data relating to the participant to be stored on different databases, but still allows accurate retrieval of information for those who have the correct access privileges.

Archiving data is not glamorous, is frequently forgotten, but is essential. Storage of raw data is critical as those data may be requested for analysis later, for example as part of a systematic review when the original study omitted some of the data in the results section.

Electronic database systems facilitate data archiving, but only when validated procedures take place according to the protocol set by the data manager. It is sensible to back-up data on at least two different locations, for example two physically separate hard drives, and also to include a back-up in a different physical location. Online back-up has progressed and is much less expensive. However, it should also be validated in terms of data control so that sensitive data are protected.

C6. Clinical data

Calibration of examiners, even if only one examiner is used, needs to be undertaken and recorded. Intra-rater reliability refers to the consistency of a single examiner to reproduce the same measurement and is achieved through training and continuous monitoring. Inter-rater reliability refers to the consistency of more than one examiner to agree and this agreement may be assessed statistically using kappa scoring (Cohen's or weighted kappa).¹⁰

There are two main methods in clinical dentistry to calibrate examiners. The first, but more difficult to set up, is that examiners are asked to examine independently a randomly selected subgroup of study participants over a short period and then compare results. The other method is to appoint a 'gold standard' (reference) examiner against whom all the examiners are checked. Comparisons among examiners are then related to this reference examiner.

It is important to collect all relevant clinical outcomes at the time of the study, although it should not be a trawling process to collect as much data as possible in case they can be used at a later date. The data-collection process will have been discussed at the outset of the research project, defined at the protocol stage, and trialled before the main study began.

D. Sampling, randomisation and validity

D1. Sampling


It is not feasible to survey an entire population to collect data, so the method used is to select a subset, or sample, of that population. The sample has to be truly representative of that population or the external validity will be poor and the results of the study will then not necessarily apply to the population as a whole. The process of sampling should be done by random selection, using one of the methods described earlier. Random selection¹¹ means that there is an equal chance that any member of a given population may be selected.

The sample size should be determined before the study starts and the process recorded in the protocol. A power calculation should be performed before the study pilot, usually with the help of a statistician to define the parameters. The calculation can be left to dedicated software as an 'a priori' power calculation. The power calculation determines the minimum number of participants and thus the data that must be collected. If the sample size is too low, then it is likely that Type II errors will result. Type II errors are false negatives and the statistical analysis of the study, for example a randomised controlled trial, will show that the intervention produced the same results as the control. However, the true effect (that is, if there was an effect) is masked because the sample size was too small. A 'post hoc' power calculation is performed after the study is completed and its value is lower, because it is too late to change the study, but it may be of value if the study is repeated.

D2. Randomisation

There are many experimental designs that can be used and the simplest form of experimental design to look for a cause–effect relationship is the Pretest–Posttest Control Group design.¹²

Randomisation of participants is carried out and assigned to either the intervention group or the control group. Both groups are observed at the start of the study; group 1 has the intervention carried out over a period of time, whereas the control group has no intervention (Table 3). At the end of the study, both groups are examined to see whether any changes have taken place. This study design is commonly used in healthcare research and answers the question about the intervention having an effect and also eliminates other explanations as far as possible, for example confounding variables. The random assignment of participants to the two groups assures that any differences present between the groups are due to chance only.

Table 3: Randomisation in a pretest–posttest control group study				
	Group	Time 		
Random assignment	Group 1	Observed	Intervention	Observed
Random assignment	Group 2	Observed	-----	Observed

There are several ways to generate randomly samples. The oldest is probably random number tables. Random number tables are better than rolling dice, and so on, and are typically presented as blocks of numbers; for example, 100 blocks of numbers with 10 rows and 10 columns of blocks. The rows and columns are numbered to help find the starting point in the table, which must itself be randomly chosen. More recently, many computer programs, for example a Microsoft Excel spreadsheet (Microsoft Corporation, Redmond, WA, USA), have a random number generator function and there are now a plethora of free online random number generators that may be used.

D3. Validity

Validity¹³ is about the accuracy and truthfulness of the research project as a whole and is used to ask whether the results of the study itself are warranted and if the results can be applied to the population as a whole. Validity is thus about the individual study and, also, whether the results of that study can be translated to the general population.

Internal validity seeks to address whether cause and effect conclusions are justified by the data. It seeks to remove all other possible explanations for the results. Good study design is crucial in ensuring good internal validity. For example, in clinical research, a double-blinded randomised controlled trial, if well conducted, will have the best internal validity for a single study design.

External validity concerns whether the results may be generally applied to the wider population. There are three methods commonly used to enable this to happen. First, the use of a real-life setting; in healthcare, this means in vivo (in life) rather than in vitro (in glass) studies. It is common for data collected and analysed for in vitro laboratory studies to be unreliable when applied to the population. Second, a representative sample; this may be

difficult given the diversity of the population. Third, replicating the study in a different context; when the researchers reach the same conclusion in all their studies can validate the conclusion when taken together.

E. Final thoughts

Data collection is crucial to a research project and failure to collect data accurately may lead to the following:

1. The research hypothesis is not answered correctly.
2. The study has problems with validity and is difficult to repeat.
3. Incorrect findings may cause harm to participants when applied in a wider setting.
4. Public health policy may be compromised as a result of misleading analyses and conclusions inferred from inaccurate data.

F. Suggested further reading

- Leedy PD, Ormrod JE. Practical Research Planning and Design. 9th ed. Upper Saddle River; NJ: Pearson; 2010.
- Giannobile WV, Burt BA, Genco RJ. Clinical Research in Oral Health. New York: Wiley-Blackwell; 2010.

References

1. Suvan J. An introduction to research for primary dental care clinicians. Part 6: Stage 7. Piloting the methodology and project management. Prim Dent Care. 2011;18:127-32.
2. Toy A, Eaton KA, Santini A. An introduction to research for primary care clinicians. Part 2: Stage 4. Planning the study. Prim Dent Care. 2011;18:36-40.
3. Giannobile WV, Burt BA, Genco RJ. Clinical Research in Oral Health. New York: Wiley-Blackwell; 2010. p. 281-2.
4. The International Caries Diagnosis and Assessment System. Accessed (2011 Sep 30) at: www.icdas.org/
5. Löe H, Silness J. Periodontal disease in pregnancy. 1. Prevalence and severity. Acta Odontol Scand. 1963;22:533-51.
6. Santini A, Eaton KA. An introduction to research for primary dental care clinicians. Part 4: stage 6a. Obtaining ethical approval. Prim Dent Care. 2011;18:127-32.
7. Kaplan C. Transforming the back office with a single keystroke. Healthc Financ Manage. 2011;6:118-20, 122, 124.

8. Whitney CW, Lind BK, Wahl PW. Quality assurance and quality control in longitudinal studies. *Epidemiol Rev.* 1998;6:2071-80.
9. Derby DC, Haan A, Wood K. Data quality assurance: an analysis of patient non-response. *Int J Health Care Qual Assur.* 2001;24:198-210.
10. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70:213-20.
11. Hinkelmann K, Kempthorne O. *Design and Analysis of Experiments: Introduction to Experimental Design.* 2nd ed. Hoboken, NJ: Wiley; 2008.
12. Campbell DT, and Stanley JC. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally; 1966.
13. Creswell JW. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* 3rd ed. Los Angeles: Sage; 2009.