IN BRIEF
- The distinction between observational and experimental studies
- The use of the relative risk in a cohort study and the odds ratio in a case-control study
- A description of parallel, matched and cross-over designs
- The choice of observational unit

Further statistics in dentistry
Part 2: Research designs 2

A. Petrie[1] J. S. Bulman[2] and J. F. Osborn[3]

To facilitate statistical discussion, it is essential to have an understanding of factors, effects, interactions, confounding, bias, estimation and hypothesis testing, to name but a few terms commonly used in statistical investigations in dentistry. These terms were explained in the previous paper in this series. This second paper is concerned with describing and differentiating between the various research designs that may be adopted.

[1]Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London; [2]Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London; [3]Professor of Epidemiological Methods, University of Rome, La Sapienza
Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD
E-mail: a.petrie@eastman.ucl.ac.uk

OBSERVATIONAL VERSUS EXPERIMENTAL DESIGNS

Studies may be either observational or experimental. An **experimental** study is one in which the investigator deliberately intervenes so that it is possible to observe the effect of the intervention on the response of interest, usually with a view to establishing whether a change in the response is attributable to the intervention. A clinical trial is an example of an experimental study. An **observational** study is one in which the investigators do not intervene in any way, so they do not, for example, administer treatments or withhold factors which may influence the outcome of interest. An **epidemiological** study is concerned with investigating the effect of certain factors and their inter-relationships on disease. The study is usually devised with a view to eliciting possible causes of the disease, and is generally observational rather than experimental because the potential aetiological factors are often not amenable to random allocation, perhaps for ethical reasons. So, for example, in a study of the effect of cigarette smoking on the incidence of oral cancer it would be impossible (illegal and unethical) to randomly allocate individuals or communities to various levels of consumption of a potential carcinogen.

Both experimental and observational studies have much in common and it is perhaps unfortunate that some people regard the methodology of experimental research as 'medical (or dental) statistics' and the methodology of observational studies as 'surveys' or 'epidemiology'.

The effects of suspected confounding variables can be investigated in an observational study. However, if confounding variables exist without being suspected, they may misleadingly distort the apparent effect of the risk factor under study. This is the main disadvantage of an observational study; the observed effect of the factor under investigation may be due to an unsuspected confounding factor.

OBSERVATIONAL STUDIES

Observational studies may be **cross-sectional** or **longitudinal**. Cross-sectional studies provide a snapshot picture of a community at a point in time, and do not involve following a group of individuals over time. In contrast, **longitudinal** studies are those which require the individuals to be investigated over a period of time. The study may be **prospective** (eg a cohort study) in which case the data are collected forward in time from a given starting point. On the other hand, **retrospective** studies (eg case-control studies) are those in which the information on the individual is obtained by going backwards in time to events that have occurred, possibly relying on case records to obtain the relevant information. It should be noted that although experimental studies, by their very nature, are invariably longitudinal, observational studies may be either cross-sectional or longitudinal.

The advantage of cross-sectional studies is

that they are fairly quick, easy and cheap to perform. However, they cannot provide evidence of a temporal relationship between the risk factors and disease since the data concerning exposure to the factor and the presence or absence of disease are collected simultaneously.

Sample surveys

A **sample survey** is a particular form of cross-sectional observational study in which a sample of individuals is taken from a well defined population with the intention of using the observed characteristics in the sample as estimates of the corresponding characteristics in the population. In general, the sample might be used to estimate any characteristic, but commonly the estimate would be the average value of some measurement such as age, systolic blood pressure etc or an estimate of the proportion of the individuals in a population who possess a particular attribute. If the attribute were a disease, this proportion would be called the prevalence of the disease.

Table 1 Frequencies of individuals with some or no tooth loss in a cohort study			
	Current smoker	Non-smoker	Total
Lost ≥ one tooth	80 (32%)	120 (16%)	200
Lost no teeth	170	630	800
Total	250	750	1000

Cohort studies

A cohort study (Fig. 1) involves observing and monitoring a group of individuals over a period of time. Such studies come under a number of different guises, particularly in epidemiology, with names such as *cohort* studies, *longitudinal* studies and *prospective* studies, and although these studies may involve different definitions of the study population, the statistical analysis of each is essentially the same.

In a cohort study, the individuals in the sample from the relevant study population are first categorised according to the levels of the factor or factors of interest, perhaps a risk factor such as daily cigarette consumption so that each individual is classified as a current smoker or non-smoker. This cohort of individuals is then monitored for a period of time and a change in status is noted. In an epidemiological study, the status may change, for example from 'without disease' to 'with disease', where the 'disease' might be oral cancer or the loss of at least one tooth. Such changes may be measured by the rate at which new cases of the disease occur in the study population. This rate is usually called the *incidence rate* of the disease. The observed incidence rates in the risk factor categories are then compared, usually by calculating their *ratio,* called a **relative risk.**

Suppose, for example, a sample comprised 1000 individuals aged 60+ years, each of whom had an oral examination and interview at baseline and then again 2 years later. Two hundred (20%) of the individuals lost one or more teeth during the 2 year period. Eighty (32%) of the 250

of the individuals categorised as current smokers at baseline and 120 (16%) of the 750 non-smokers lost at least one tooth in that time (Table 1). The relative risk of tooth loss is thus estimated as RR = 32/16 = 2.0, indicating that the risk of tooth loss in current smokers was twice that of non-smokers. A relative risk of one implies that the risks of disease in those exposed to the factor and those not exposed are the same. A relative risk greater (or less) than one shows the extent to which the risk of the disease in the exposed group is increased (or decreased) relative to that of the unexposed group. The confidence interval for the true relative risk is evaluated by first determining the standard error (SE) of the $\log_e$ of the relative risk, and then using the theory of the Normal distribution.[1] In particular, $SE(\log_e RR) = \sqrt{\{1/80 - 1/250 + 1/120 - 1/750\}} = 0.124$ and the 95% confidence interval for the RR in the tooth loss example is $\exp\{\log_e 2 \pm 1.96 \times 0.124\} = 1.57$ to 2.55. This interval does not contain one and so there is evidence ($P < 0.05$) that the risk of tooth loss is significantly greater in current smokers than in non-smokers.

Sometimes, particularly for policy development, it is useful to measure how much disease burden is caused by certain modifiable risk factors. For example, the investigator may wish to answer the question 'Amongst smokers, what percentage of the total risk of tooth loss is due to smoking?' A suitable measure that provides an answer to this question involves the calculation of the **attributable risk** which is the difference between the tooth loss incidence rates in the risk factor categories.

Although a cohort study is time-consuming and costly, and is useful only for studying a common disease, it has the advantages that it can be used to study many disease outcomes as well as rare risk factors.

Case–control studies

In a case-control study (Fig. 1), sometimes called a c*ase-referent, retrospective* or *trohoc* (cohort spelt backwards) study, a sample of cases, ie persons diagnosed as having the disease of interest, is compared with a group of comparable controls who do not have the disease. The cases and controls are separately categorised according to whether or not each has been exposed to the risk factor. Since it impossible to estimate the relative risk directly in a case-control study (as the relative risk requires knowledge of disease rates rather then exposure rates), it is common to estimate the **odds ratio** instead. The odds of the disease in those exposed to the factor is the chance of having the disease in those exposed to the factor divided by the chance of not having the disease in this group of individuals. The odds of disease in those not exposed to the factor is defined in a similar fashion. Then the odds ratio is the odds of disease in those exposed to the factor divided by the odds of disease in those not exposed to the factor. The odds ratio is a reasonable estimate of the relative risk of disease in those who are and are not exposed to the factor

provided the disease is rare and so its prevalence is low.

Consider, for example, a case-control study which was performed to investigate the association, if any, between betel nut chewing and oral mucosal lichen lesions in women in Cambodia.[2] It was found that 5 (23.8%) of the 21 women with lichen lesions chewed betel nut, while among the 1,469 controls (ie women without lichen lesions), 127 (8.6%) chewed betel nut (Table 2). So the estimated odds of lichen lesions in those who chewed betel nut was (5/132)/(127/132) = 5/127, and the estimated odds of lichen lesions in those who did not chew betel nut was (16/1358)/(1342/1358) = 16/1342. The prevalence of lichen lesions in this group of women was low and equal to 100 x 21/1490 = 1.4%. Hence, the estimated odds ratio of (5/127)/(16/1342) = (5 x 1342)/(127 x 16) = 3.3 could be used to estimate the relative risk. This implies that the risk of lichen lesions was 3.3 times greater in women who chewed betel nut than in those who did not chew betel nut. A confidence interval can be determined for the true odds ratio since it can be shown that the sampling distribution of $\log_e(OR)$ approximates a Normal distribution and that $SE[\log_e(OR)] = \sqrt{(1/a + 1/b + 1/c + 1/d)}$ where a, b, c and d are the numbers of individuals exposed and not exposed to the risk factor in those with and in those without the disease. In the lichen lesion example, $\log_e(OR) = 1.19$ and $SE[\log_e(OR)] = \sqrt{(1/5 + 1/16 + 1/127 + 1/1342)} = 0.52$. Thus the 95% confidence interval for the logarithm of the true odds ratio is $\log_e(OR) \pm 1.96 \times SE[\log_e(OR)] = 1.19 \pm 1.96 \times 0.52 = 0.173$ to 2.215. Hence the 95% confidence interval for the true odds ratio is $e^{0.173}$ to $e^{2.215} = 1.19$ to 9.16. This confidence interval excludes one indicating that the odds ratio is significantly different from one ($P < 0.05$) and that the risk of lichen lesions in the Cambodian women from which this sample was taken was significantly greater if they chewed betel nut.

This essentially simple design can be elaborated to include stratification, matching and regression analysis to control the influence of confounding variables on the estimated relative risk. Multiple regression is discussed in greater detail in a later paper in this series.

The disadvantages of a case-control study are that it is not possible to estimate the relative risk directly from the study (although if the prevalence of the disease is low, the odds ratio can be used as an estimate of the relative risk), that selection of the controls may be difficult and that it is possible to study only a single disease outcome in any one study. However, case-control studies are relatively quick, easy and cheap to perform, and can be used to study many risk factors as well as rare diseases.

EXPERIMENTAL STUDIES

If the study is experimental rather than observational then it must be designed in such a way that it gains the largest amount of information
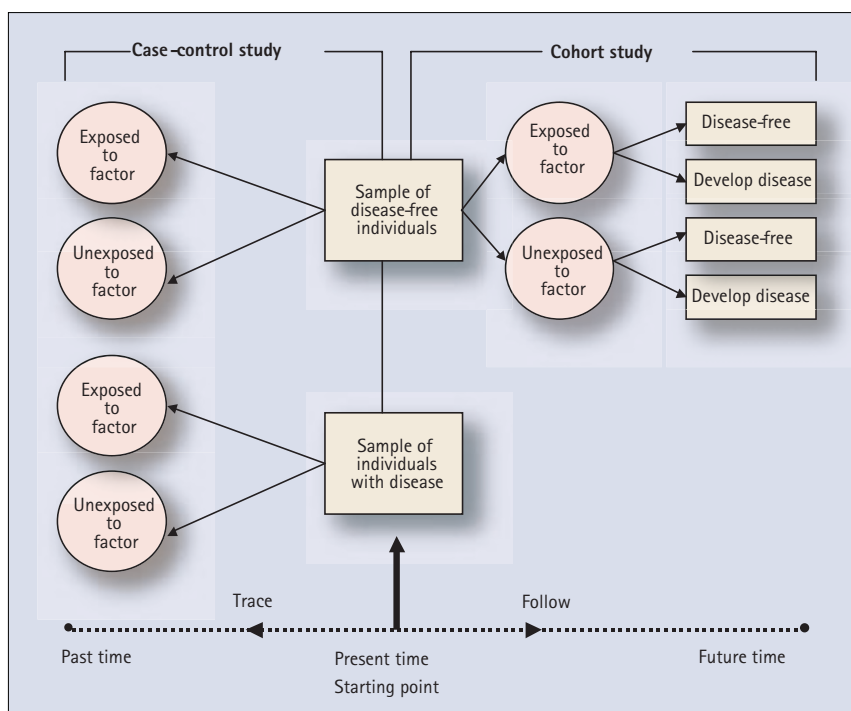

Fig. 1 Diagrammatic representation of a case-control and a cohort study

Table 2 Frequencies of women with and without lichen lesions in a case-control study

	Women with lichen lesions	Women without lichen lesions	Total
Chewed betel nut	5 (23.8%)	127 (8.6%)	132
Did not chew betel nut	16	1342	1358
Total	21	1469	1490

of the greatest reliability in an efficient manner. The objective, therefore, is to achieve an optimal balance between minimal sample size and maximum precision whilst eliminating sources of bias and identifying and controlling all sources of variation. This balance may be achieved by choosing the appropriate experimental design which takes into account the particular circumstances of the investigation.

Invariably, a well-designed experiment is both **comparative** and **randomised.** The comparison is usually between the unauthenticated novel intervention (such as a treatment or preventative measure) and some form of 'control', such as an established intervention. Randomisation, also called random allocation, implies that the subjects are randomly (ie using a method based on chance) assigned the treatments or interventions. One advantage of randomisation is that potential confounding factors will be approximately evenly distributed in the different intervention groups. So, for example, in a study of the effects of a therapeutic dentifrice in the treatment of periodontal conditions in a large multiracial society, random allocation of the subjects to the dentifrice or control 'treatments' would ensure that each ethnic group is approximately equally represented in both the study and

The clinical trial

The clinical trial is a particular form of experimental study

control groups. This would be important if ethnic group were associated both with the use of the dentifrice and the periodontal condition, with consequent difficulties in separating the effects of these factors on the outcome.

The **clinical trial**[3] is a particular form of experimental study which is afforded special consideration because the experiment is performed on humans. Particular attention must be focused on the ethical problems that arise in medical and dental research. Designing the trial so as to use the minimum number of patients enabling a valid conclusion regarding the efficacy of treatments to be drawn must be a major objective in the clinical scenario. A full discussion of the clinical trial, randomisation and sample size calculations will be given in two later papers.

One important distinguishing feature of any experimental design is whether the treatment comparisons are made *between* subjects (parallel groups designs) or within subjects (matched designs or cross-over studies).

Parallel groups

Parallel groups designs involve the basic observational units (typically, the subjects) being independently and randomly allocated to two or more treatment groups. The response is observed for every individual in the study and an aggregate measure (usually an arithmetic mean or median if the response is quantitative or a proportion if the response is qualitative) is calculated for each treatment group. These summary measures are then compared appropriately so that the investigator can determine whether the responses differ significantly in the different treatment groups. The parallel group design therefore relies on comparisons which are made between groups of subjects. It should be noted that although generally desirable, it is not necessary to have an equal number of subjects in each group.

If there are two treatment groups and the response is quantitative and satisfies the assumptions underlying the method, the comparison of response to treatments may be afforded by the two-sample *t*-test. If there are more than two treatment groups, the one-way analysis of variance facilitates treatment comparisons, provided the assumptions underlying the method are satisfied. If the response is qualitative, the Chi-square test is often employed for comparative purposes.

The randomised parallel groups design has the advantages that it is conceptually simple and the analysis is straightforward. In some circumstances, however, it may be appropriate to modify the simple parallel group design by employing a technique called **blocking** or **stratification** in addition to the simple randomisation of subjects to treatments. This involves forming subgroups of individuals, the blocks or strata, such that the variation with respect to the variable of interest within each stratum is smaller than the variation between the strata. Consider, for example, an analysis of the variable DMF which

is higher in older children than in younger children. It may therefore lead to greater precision for a given total sample size (or alternatively equal precision for a smaller sample size) if the overall group of children is stratified by age, and the older age-group analysed separately from the younger. In other words, the individuals are randomly allocated to the different treatments in each age stratum so that a simple parallel groups design is contained within each of these age strata. Subsequent treatment comparisons are made between groups of subjects within each stratum, and the results properly combined to determine the overall treatment effect.

Stratification may also be employed because it is of interest to investigate whether the effect of treatment (say the difference in response in the two or more treatment groups) is the same for all strata of the study population. For example, is the effect of treatment the same for younger children as it is for older children? If the treatment effect depends on the factor defining the blocks or strata, there is an interaction between the treatment and the factor. This would clearly be important for identifying patients who would benefit from a new treatment.

Even if the effect of the treatment or intervention were the same at every level of the blocked or stratified factor, the response might change systematically with the factor. For example, the average effect of the treatment (that is, the difference between the average responses to two treatments), may be the same in every age group, but the response may tend to increase with age. By making the comparison between the two treatment groups within each age group, the factor age will not confound the treatment effect. Furthermore, by controlling the potential confounding effect of a variable such as age, the precision of the comparison between the two groups will be improved.

Thus the advantages of blocking or stratifying the study population before randomisation are to enable interactions to be detected and estimated, to control the effect of known potential confounding factors and to improve precision. The disadvantage is that the statistical analysis is slightly more complicated.

Matched designs

If the blocking described above is carried to extremes, then pairs of subjects (or triplets if there are three treatment categories) can be matched so that they are alike with respect to a number of potential confounding factors. For example, if it were decided to match for age and sex, the subjects in the study would be arranged in pairs so that the two individuals in each pair would be the same age and sex. The two individual subjects in each matched pair would then be randomly allocated to different treatment/intervention groups. The comparison between the two treatments is made within each matched pair and thus the treatment effect will be more precisely estimated than it would be with a parallel groups study with the same number of subjects.

Between– and within–group comparisons

Comparisons are made:
• Between groups in a parallel groups design
• Within groups in matched pair and cross–over designs

The analysis of matched studies is relatively straightforward and is often achieved by using the paired *t*-test for matched quantitative data or, if the data are dichotomous, McNemar's test.

The advantage of a matched study compared with a parallel groups design is a gain in precision with the same number of subjects, or equivalently, the same degree of precision of a parallel groups study can be achieved with a smaller total number of subjects.

The disadvantages of matching are that the study may become logistically difficult if too many matching factors are included and the inability to match some subjects may reduce the total number of subjects in the study. It may be more difficult to investigate interactions in a matched study.

Cross–over trials

The matched pairs study enables treatment comparisons to be made using similar experimental units. Rather than these experimental units being different subjects who have been matched appropriately, a similar type of study is one in which the subject acts as his/her own control with the same subject being allocated both treatments, receiving them at different times. Such designs are called cross-over designs[4] because the subject crosses over from one treatment to the other. The designs should involve randomising the order of administration of the treatments to each subject. The treatment comparison is then made within subjects and, in the same way as a matched pairs study, increases the precision of the treatment effect for a given number of subjects.

Designs in which the subject receives both treatments are sometimes regarded as an extreme form of matching. However, the difference between extreme matching and using the subject as her/his own control arises because with matched pairs the subjects are randomly allocated to treatments, whereas in the cross-over trial the subject acts as his own control and thus receives both treatments. In the analysis of simple studies, this difference may not matter but with more complicated designs the fact that the main observational unit, the subject, is split between the two treatments may need to be taken into account.

Cross-over trials, although advantageous when compared to parallel groups designs in terms of precision or sample size, cannot be utilized for conditions which do not remain stable in the study period or which can be cured by the treatments being administered, when there is a carry-over effect from one treatment to another, or when the response to treatment is prolonged.

The choice of observational unit

A fundamental consideration in research designs concerns the choice of observational unit.[5] It is important to understand that the unit of observation in an experiment or observational study is the smallest unit with a unique set of important characteristics which is independent of other similar units in that its response cannot be affected by these other units, and which can be assigned to each of the treatments in an experimental study. Thus the observational or experimental unit in a clinical trial is often the patient or, in the case of dental investigations, the mouth because teeth cannot be regarded as independent units within the mouth. The experiment should be designed and analysed with this in mind so, for example, the randomisation process should randomise the mouths (the experimental units) rather than the teeth (the sub-units) to the different treatments. In the same way, the sample size estimation process whilst satisfying certain criteria, must aim to estimate the optimal number of experimental units rather than the sub-units contained within them.[6]

As an example, consider just two situations where either the individual child or a 'community' of children, say a school, is the basic unit of observation. For example, in a randomised intervention study of fluoride supplement, if the individual child was the basic unit, individual children would be randomly allocated to receive the intervention or not, whereas if the basic unit was the school, then the schools would be randomly allocated and the responses would be observed for individual children within their school. The difference between these two types of design is very important. An extreme example may make this clearer. Suppose 1,000 children attend ten schools and it is of interest to investigate the effect of fluoride supplementation on DMF. Two designs that might be considered are:

1. Take each child and randomly allocate it either to receive the supplement, or not, and after 1 year compare the means of the changes in DMF in the two groups of 500 children.
2. Give the supplement to all the children in five randomly chosen schools and withhold it from all the children in the other schools. Calculate the mean change in DMF in each school and then compare the means of these mean changes in the two groups of five schools.

Clearly, in Design 1, where the individual child is the basic unit, a more precise estimate of the effect of supplementation (ie one with a narrower confidence interval) will be obtained than in Design 2 where the comparison may be confounded by other differences between the schools. Design 2 could be improved if there were many more schools available for randomisation.

The advantages and disadvantages of the two designs are:

1. For a given total number of children, studies with the child as the basic unit will be more precise and have a greater power to detect an effect of treatment than if the schools are the observational units.
2. Equivalently, to achieve a given level of precision, more children are required for a school based study than if the children themselves are the observational units. This increase in

The observational unit

It is more usual to take the mouth rather than the tooth as the unit of observation in a dental investigation

sample size (or loss of precision) is often called the *design effect* of the study.

3. Logistically, it is often much easier to organise and administer a study based on schools rather than children. In extreme cases, for example a community intervention such as the introduction of piped water, it may be impossible to conduct a study based on individual persons.

4. For a given total monetary budget (including the costs of all resources used, such as manpower, equipment, travel etc), it will usually be possible to have a larger total sample size if schools are the observational units. This increase in sample size will sometimes more than offset the advantages, discussed in Points 1 and 2 above, of studies with child based units

5. Analysis is usually easier if a study is based on independent individuals rather than schools, but clearly the availability of computers and user-friendly packages for statistical analysis makes this advantage less important.

In a sample survey, the simplest design in which the observational unit (for example, a village) comprises a collection of individual units (for example, people) leads to a *cluster* analysis.[7] The clusters (the villages) are randomly selected and all the individual units (people) within each selected cluster are observed. This design may be extended to multi-stage or hierarchical sampling.

In an experimental study, the design in which the main experimental units (for example, mouths) containing sub-units (for example, teeth) are assigned to different treatments leads to a *split-plot, split-unit, nested*[8], *multi-level* or *hierarchical*[9] analysis. The difficulty with analysing such designs is that there are two

sources of sampling error: that arising from the differences between sub-units units within each main unit and that caused by differences between main units. In almost all situations, the contribution of the differences between main units to the overall sampling error will be much greater than that contributed by sub-units within each main unit. It can be shown that for a fixed total study size, it is desirable (but more costly) to have a large number of main units and to observe fewer sub-units in each main unit. This same problem arises in clinical trials in which repeated observations are made on each subject. An example of such a clinical trial is a study of gingivitis in which there are three treatments, a variable number of patients in each treatment group and a variable number of sites where the gums are inflamed within each patient's mouth. The main units are the patients and the sub-units are the sites. Some aspects of the problem of the choice of units to use for the statistical analysis are considered in a subsequent paper on repeated measures.

1 Altman D G. *Practical Statistics for Medical Research.* UK: CRC press, 1991.
2 Ikeda N, Handa Y, Khim S P, Durward C, Axell T, Mizuna T, Fukano H, Kawai T. Prevalence study of oral mucosal lesions in a selected Cambodian population. *Community Dent Oral Epidemiol* 1995; **23:** 49-54.
3 Pocock S J. *Clinical Trials: a Practical Approach.* Chichester: Wiley, 1983 .
4 Senn S. *Cross-Over Trials in Clinical Research.* Chichester: Wiley, 1993.
5 Altman D G, Bland J M. Units of analysis. *Br Med J* 1997; **314:** 1874.
6 Kerry S M, Bland J M. Sample size in cluster randomisation. *Br Med J* 1998; **316:** 549.
7 Cochran W G. *Sampling Techniques.* 3rd edn. Wiley: New York, 1977.
8 Scheffé H. *The Analysis of Variance.* New York: Wiley, 1999.
9 Sullivan L S, Dukes K A, Losina E. An introduction to hierarchical linear modelling. *Stat Med* 1999;**18:** 855-888.