Design of Studies for Medical Research

DAVID MACHIN

Division of Clinical Trials and Epidemiological Sciences,
National Cancer Centre, Singapore,
UK Children's Cancer Study Group, University of Leicester, UK
Institute of General Practice and Primary Care, School of Health and Related Sciences,
University of Sheffield, UK

and

MICHAEL J. CAMPBELL

Medical Statistics Group, Institute of General Practice and Primary Care, School of Health and Related Sciences, University of Sheffield, UK

John Wiley & Sons, Ltd

Design of Studies for Medical Research

Design of Studies for Medical Research

DAVID MACHIN

Division of Clinical Trials and Epidemiological Sciences,
National Cancer Centre, Singapore,
UK Children's Cancer Study Group, University of Leicester, UK
Institute of General Practice and Primary Care, School of Health and Related Sciences,
University of Sheffield, UK

and

MICHAEL J. CAMPBELL

Medical Statistics Group, Institute of General Practice and Primary Care, School of Health and Related Sciences, University of Sheffield, UK

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Machin, David.
Design of studies for medical research / David Machin, Michael J. Campbell.
p. cm.
Includes bibliographical references and index.
ISBN 0-470-84495-7 (alk. paper)
1. Medicine–Research–Methodology. I. Campbell, Michael J. II. Title.

R850.M23 2005
610′.72–dc22

2004065408

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 470 84495 7 (PB)

Typeset by Dobbie Typesetting Ltd, Tavistock, Devon
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wilts
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

To
Christine Machin
and
Jacinta Campbell

Contents

Preface

There are many textbooks on medical statistics, but the majority concentrate on statistical analysis. However, unless care is taken as to how the data were collected in the first place, no amount of sophisticated analysis can save the experimenter from possibly making misleading conclusions. A poorly designed study is like a house built on sand, easily washed away when the design flaws are pointed out. It appears to us that few textbooks place sufficient emphasis on design of studies and so the purpose of this book is an attempt to fill this gap.

In general design books concentrate on the design of experiments. We have broadened this to include chapters on the design of surveys, epidemiological studies and studies concerned with diagnosis and of prognostic factors. Emphasis is also placed on estimating an appropriate study size and how to choose subjects for inclusion in a study. Much data are captured on forms or questionnaires and since we feel this area to be somewhat neglected by statisticians, we have included a chapter covering it. Although it may not appear to be of immediate relevance to good design, we also cover the essential care to be taken when describing the study design in any eventual publication.

Our plan with this book is to emphasise the importance of good design, whether in preclinical or clinical studies, clinical trials or epidemiological research. We concentrate on research of all types involving human subjects, although many of the designs considered are applicable to laboratory bench and animal studies. We have purposely avoided giving details of statistical analysis, although some of these are unavoidable.

We hope this book will prove useful to investigators with the design of their studies, when completing a research proposal or ethics form, and also for those doing a research methods course.

We would like to thank colleagues in Leicester, Sheffield and Southampton, UK, in Singapore and in Skövde, Sweden, for encouragement and advice. We would also like to thank colleagues, and students for bringing their design problems to us.

David Machin
Michael J. Campbell

Southover, Dorset and Sheffield

July 2004

1 What is Evidence?

Summary

This chapter introduces the ideas associated with evidence-based health care and contrasts this approach with earlier approaches in clinical medicine which had largely relied on describing pathophysiological processes. We consider the nature of proof using evidence and describe the Bradford-Hill criteria which are useful in determining how reliably causation has been established in a study. We define broad areas that distinguish laboratory and animal experiments from studies and clinical trials in humans. Experimental design can be appropriate to research in preclinical, clinical and epidemiological studies. Statistical models are at the heart of the design of studies and the purpose of a good design is to estimate the parameters of a model as efficiently as possible.

We also emphasise the need to check local regulations with respect to ethical clearance of studies and informed consent from the study participants. It is important to develop a formal protocol for any study and describe in general terms the contents of such a protocol. Published guidelines and standards for reporting the results of studies are useful pointers for consideration by the study design teams.

1.1 INTRODUCTION

This is the era of 'evidence-based medicine' (EBM) or more comprehensively 'evidence-based health care' (EBHC). EBM requires that we should consider critically all evidence about whether, for example, a treatment works, an agent causes a disease, or a drug is toxic. This requires a systematic assembly of all available evidence followed by a critical appraisal of this evidence. Before this paradigm had been formulated, biomedical investigators considered it sufficient to understand the pathophysiological process of a disorder. As a consequence the physician would prescribe to patients with relevant symptoms drugs, or other treatments, that had been shown to interrupt this process. Thus the practice of medicine had been based on history taking and clinical examination followed by treatment of symptoms, all based on the accepted pathophysiology of the condition diagnosed at the relevant time.

Design of Studies for Medical Research. D. Machin and M. J. Campbell
 ISBN 0 470 84495 7

Example *– removing the cause – ventricular ectopic beats*

Sackett, Richardson, Rosenberg and Haynes (1997) give an example of a finding that patients who displayed ventricular ectopic beats after a myocardial infarction had occurred were at high risk of sudden death. Following this observation drugs were then widely prescribed to suppress these ectopic beats, on the assumption that removing the cause would reduce the effect. However, subsequent randomised controlled trials which examined clinical outcomes, and not the physiological process alone, showed that use of these drugs actually *increased* death rates rather than decreased them. The use of these drugs is now contra-indicated.

Example *– reducing the risk – premature babies*

Gilman, Cheng, Winter and Scragg (1995) describe a study related to concerns of neonatologists who had always kept premature babies lying on their fronts. One tacit assumption was that, should the premature baby vomit, the baby would be less likely to inhale the vomit. This practice was extended to all babies. However, subsequent epidemiological studies showed that babies who were habitually put on their fronts were exposed to a higher risk of sudden infant death. A 'back-to-sleep' campaign was initiated and the sudden infant deaths in England and Wales dropped from some 2000 to less than 600 per year as a direct consequence. The argument for putting babies to rest on their fronts, albeit reasonable in nature, was not *evidence-based.*

Systematic reviews combine the evidence from individual studies to give a more powerful analysis of any effect. It is important to realise that they can only be as good as their component parts. Thus if the studies being reviewed are of poor quality then inferences drawn from an overview will have to be made with extreme caution. In contrast, if the basic information is of high quality then their collective and systematic review and synthesis clearly adds substantially to the evidence base for clinical medicine.

1.2 EVIDENCE AND PROOF

Any discussion of EBM gives rise to the question, what is evidence? The first concern is with the problem of *proof* and philosophers have long argued over this. In mathematics, the ancient Greeks demonstrated rigorous proofs of many *theorems* (literally God-like things), especially in algebra and geometry, and they thought of these as general laws.

Thus, we know for certain that Pythagoras' Theorem is true. The question arises as to whether one can have similar certainty in other areas of human enquiry.

In the natural sciences, Francis Bacon (1561–1626) described the work of scientists as collecting information and adducing natural laws. However, David Hume (1711–1776) concluded that no number of singular observations, however large, could logically entail an unrestricted general statement. Just because event A follows event B on one occasion, it does not follow that event B will be observed the next time we see A. Thus it does not logically follow, in the manner that a mathematical theorem is true, that A will always follow B whether we observe A and B together on two, twenty or two thousand occasions. The point here is that simply *observing* an association is not proof that an association actually exists.

There may, however, be real reasons why two events are associated, and in general one would hope to discover these. Thus, although we observe that 20 consecutive bed-ridden patients develop pressure sores, this does not logically imply that the 21st patient will do so. However, it does suggest a pattern that would be foolish to ignore when considering appropriate care for patient 21.

'Hume's problem' troubled philosophers as it seemed to discourage endeavours to make sense of nature. It was not until the last century that Karl Popper (1902–1994) proposed the idea of falsifiability. Falsifiability states that laws cannot be shown to be either true or false but that they can only be held *provisionally* true. He pointed out that observations cannot be used to prove laws, but can falsify them. Hume's famous example is the universal law 'all swans are white'. This cannot be proven, no matter how many swans one sees that are white, but it would take only a single black swan to refute the law. This has direct bearing on statistical inference, where, as part of the study design, one sets up a null hypothesis and then tries to refute it with the experimental observations. Failure to reject the null hypothesis does not logically imply that one should accept it, rather it implies that we do not have enough evidence to reject it.

Clinical trials which compare treatments are designed with a null hypothesis in mind, namely that the treatments have no differential effect on patient outcome. We try and disprove this null hypothesis using patient data. However, we can *never* prove a null effect.

The basis of EBM is that any guidance arising from any review of evidence is only *provisional*, albeit based on the best evidence available at the time. We can collect more evidence and, if this concurs with the existing evidence, it may give us greater confidence in our guidelines, but still cannot prove them. However, later evidence may contradict the existing theories (and hence disprove them), however well founded the past evidence is.

This approach may seem rather negative, but in fact it is liberating. What Popper's philosophy gives scientists is the freedom of 'trying their best'. With this they avoid claiming omnipotence, such as would be implied if their statements were assumed true for all time. It gives scientists a model whereby criticism of existing models is actively encouraged. It enables us to differentiate the good scientific theories from the poor. For good ones, one can devise experiments to attempt to falsify the hypotheses arising from the theories. However, all theories are not equally valid. Thus theories that have withstood attempts to disprove them are to be preferred over those that have not been so tested. It is worth pointing out, however, that often the choice of *which* experiments

Table 1.1 The Bradford-Hill criteria to assess causality (after Hill, 1965; reproduced by permission of the Royal Society of Medicine)

1.	Temporality
2.	Consistency
3.	Coherence
4.	Strength of association
5.	Biological gradient
6.	Specificity
7.	Plausibility
8.	Freedom from, or control of, confounding and bias
9.	Analogous results found elsewhere

to conduct are financial, social or political decisions. Thus lack of supporting evidence for a theory may not necessarily be a deficiency of the theory itself, but rather the lack of will to *test* the theory.

Outside of the realm of mathematics, and in the less predictable fields of the biomedical and clinical sciences, the nature of human variability has meant that universal laws are rare. There are some obvious laws, such as if a person is deprived of oxygen they soon die; but such laws are the exception. Thus if we give a person a large dose of arsenic, they do not inevitably die. Rather than with establishing universal laws, biomedical science is concerned with a number of basic questions such as: Does exposure to substance *A* increase the risk of disease *B*? Does treatment *C* cure more people with disease *D* than other therapies?

More than a century ago Robert Koch (1843–1910) devised a number of questions the answers to which could be used to try and decide whether a specific bacterium caused a particular disease. These were modified by Bradford Hill (Hill, 1965) to a general examination of whether an event, such as an environmental exposure or smoking, would increase the risk of disease or prescribing a medical treatment improves the chance of cure. The Bradford-Hill criteria are summarised in Table 1.1.

In the Bradford-Hill criteria temporality means that the effect follows the cause and not vice versa. Thus a fall in lung cancer deaths in UK men succeeded a drop in the numbers of male smokers with a lag in time of some 30 years. This lag lends weight to a causal link between smoking and lung cancer. Consistency implies that the same fall in lung cancer deaths has been observed in women, or in other countries where smoking prevalence has fallen. Coherence means that different study types, such as case–control and cohort studies addressing the same issue, lead to similar conclusions. Strength of the association suggests that the stronger the effect the more plausible the causality. For example, smokers have 10 times the risk of lung cancer compared with non-smokers. The idea concerning the biological gradient is that if heavy smokers are found to be at greater risk of lung cancer than light smokers, then the case for causality is strengthened.

Specificity suggests that if the link were causal, the smokers would be mainly at risk from respiratory disease mortality, and not from other unrelated types of mortality such as those arising from road accidents. The relationship appears plausible as cigarette smoke is inhaled into the lungs and autopsy evidence from smokers and non-smokers documents clear differences between their respective lungs. A confounding

variable is one that is related to both the exposure and the outcome, but not through a causal pathway. For smoking, genetics has been argued as a confounder on the basis that the impulse to smoke may be genetic – certainly if parents smoke then children are more likely to smoke. Also genes may control the risk of lung cancer. If the genes for smoking and lung cancer were linked then it would appear that smoking and lung cancer were causally related. However, if the genetic theory were true, it would have a hard time to explain away the other causal evidence such as that provided by temporality. Bias could occur in a study or survey because people with lung cancer may be more likely to recall details of their smoking history than people without lung cancer.

Just as in philosophy we cannot prove a universal law, so in medicine we cannot prove absolutely a causal effect. Satisfying the Bradford-Hill criteria increases the likelihood that a causal effect is present, but cannot give an absolute proof of it. Hill (1965) himself admitted: 'none of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be regarded as a sine qua non'.

As one example, this philosophy has considerable implications when epidemiologists try to show that the measles, mumps and rubella (MMR) vaccine does not cause autism. We can never prove the null that that there is no association between the MMR vaccine and autism. All we can do is demonstrate that, if there is a risk, then the risk is very low. It is up to those who advise on public health issues to decide whether the risk of autism is lower and/or less damaging than the competing risks associated with a child having measles. In this respect, temporality was a major issue as in the UK increases in the diagnosis of autism had been linked to the introduction of MMR. However, this increase has not been observed in other countries, none of the other Bradford-Hill criteria are satisfied and there is no clear biological theory linking vaccines to autism.

1.3 COLLECTING THE EVIDENCE

In certain circumstances, evidence for a particular theory may be built up by a series of well-conducted experiments under very controlled (perhaps laboratory) conditions. In contrast, other information may only be obtained incidentally, such as long-term information collected from survivors of the nuclear bombs exploded in the 1940s or by the radiation leakage from Chernobyl nuclear reactors in the 1980s. Thus, it is convenient to distinguish studies in which the investigator conducting *experiments* has total control over the structure of the study and the variables to be some of the observed, and *observational* studies in which the investigator cannot manipulate the values of the variables but merely observe their value.

Control of the 'experiment' is clearly a desirable feature – perhaps easy to attain in the chemistry laboratory but not so easy with living material, particularly if they are animal or human. However, the additional difficulties imposed on the design of studies in human subjects imply that special care should be taken in the design of the studies planned. A good study should answer the questions posed as efficiently as possible. In round terms, this implies with as few subjects as is reasonably possible for a reliable answer to be obtained.

Although 'good science' may lead to an optimal choice of design, the exigencies of 'real life' may cause these ideals to be modified. Nevertheless we can still have some

Table 1.2 Characteristics of laboratory, laboratory animal and human experimental studies

Design feature	Laboratory	Animal	Human
Method of assessments	No restriction	If invasive, may not be acceptable	If invasive, may not be acceptable
Treatment or intervention	No restriction on choice of treatments – other than scientific judgement	Some procedures may bring unacceptable suffering	Implicit that treatments should do some good – thus an innocuous or placebo treatment may not be acceptable
Subject safety issues	None	Minor	Paramount – overriding principle is the safety of the subjects
Protocol review	Scientific only	Scientific and ethical	Scientific and ethical
Consent	None	None	Fully informed consent mandatory
Recruitment	Experiment can be conducted at one calendar time point	Experiment can be conducted at one calendar time point	Usually, subjects recruited one-by-one over calendar time
Time scale	May be relatively short – hours, days or weeks	May be relatively short – days, weeks or months	May be relatively long – weeks, months or years
Study size	All observations planned are made	All observations planned are made	Subjects may refuse to continue in the study at any stage
Observations	Assessed at one calendar time point	Assessed at one calendar time point	Usually, subjects assessed one-by-one over calendar time
Design changes	Immediate	Possibly ethical constraints	Almost certainly requires new ethical approval
Data protection	None	None	Confidentiality and often National Guidelines for storage
Reporting	No formal rules – journal editor's prerogative	No formal rules – journal editor's prerogative	CONSORT for Phase III trials (Begg *et al.*, 1996)

hierarchy in the choice of designs, but where we can enter this hierarchy will depend on circumstance. Thus we do not aim for the 'best' design only the 'best realisable' design in our context.

Table 1.2 illustrates some aspects of the differences that need to be considered when comparing (bench) laboratory-based (non-animal or -human) studies with clinical studies. In some sense the laboratory provides, at least in theory, the greatest flexibility in terms of the experimental design and studies in human subjects should be designed

(whenever possible) to be as close to these standards as possible. In general it can be seen that the requirements for human studies are more restrictive. For example safety, in terms of the welfare of the experimental units concerned, is of overriding concern in clinical studies, possibly of little relevance in animal studies and of no relevance to laboratory studies. As a further example, no consent procedures are required for laboratory or animal studies whereas this is a very important consideration in all human experimentation, even in a clinical trial with therapeutic intent.

1.4 TYPES OF STUDY AND HIERARCHY OF DESIGNS

For the purposes of this book we consider three broad areas of medical research. 'Preclinical' studies that are essentially laboratory-based studies and may involve human specimens or directly the humans themselves. These tend to be relatively small and afford a high degree of control for the experimenter. Examples might be studies of changes in brain image after a mental calculation or the elicitation of symptoms in a healthy person by inducing a drop in their blood glucose levels. On the other hand, 'clinical' studies are ones that involve actively intervening in the management of patients in some way, such as in a trial of a new drug. Finally 'epidemiological' studies, including surveys, broadly speaking, do not involve active intervention, but rather observe outcomes to evaluate, for example, a potential risk. Table 1.3 describes a broad

Table 1.3 The relative strength of evidence obtained from alternative designs for preclinical, clinical and epidemiological studies

	Evidence level	Type of study
Preclinical	Strongest	Blinded randomised comparative study
		Non-randomised comparative study
		Before-and-after design
	Weakest	Case-series
Clinical	Strongest	Double-blind randomised controlled trial (RCT)
		Single-blind RCT
		Community intervention study: cluster design
		Non-blinded RCT
		Non-randomised prospective study
		Non-randomised retrospective study
		Before-and-after design (historical control)
	Weakest	Case-series
Epidemiological	Strongest	Cohort study
		Case–control study
		Cross-sectional survey
	Weakest	Case-series

'hierarchy' of designs that give an increasing weight to evidence obtained from these three different types of clinical study.

PRECLINICAL

The design that can provide the strongest evidence is the randomised comparative study in which the experimental units are allocated to an intervention by some form of random mechanism as is described in Chapter 4. In a comparison between two interventions, or an intervention and a control, it is sometimes possible to give the experimental unit both interventions. In that case it is important to randomise the *order* of the interventions. A further refinement is to blind (or mask) the experimenter as to which intervention has been given to which unit. In practice, this can only be done when there are several investigators involved each with different roles in the experimental process, as another desirable feature is that the investigator doing the evaluations is also blind to the intervention received. The measures used for evaluation should also be as objective as is possible in the circumstance. Such a design is termed a *double-blind* (or double-masked) randomised controlled study. There are clearly extensions to this since one could also blind the data analyst. The purpose of the 'blinding' is to make all aspects of the study conduct to be as objective as possible and hence as free as possible from bias.

The weakest level of evidence is provided by a *case-series* that, at one extreme, may be an observation from a single unit.

CLINICAL

In parallel with preclinical studies, the design that provides the strongest type of evidence is again the *double-blind randomised controlled trial* (RCT). In this, the patients are allocated to treatment at random. In this way we can ensure that *in the long run* patients, before treatment commences, will be comparable in the intervention and control groups. Clearly, if one knew which were the important prognostic factors, one could match the patients in the intervention and control groups by other means. However, the advantage that randomisation retains is that it provides for *unknown* as well as the *known* prognostic factors, which could not be achieved by matching. Thus the reason for the intellectual attraction of the double-blind RCT is that it is the *only* design that can give us an absolute certainty that there is no bias in favour of one group compared to another at the start of the trial.

When testing new therapies, we might try a 'before-and-after' design in which outcomes before and after the introduction of the new therapy are compared. This is a very plausible scenario. After all, Alexander Fleming (1881–1955) did not need a clinical trial to demonstrate the efficacy of penicillin. Before penicillin became available most people with certain bacterial infections died, afterwards they survived. The main disadvantage of 'before-and-after' designs is that we have no idea whether the patients in the 'before' group and those in the 'after' group are comparable. Whilst it is hard to imagine the natural history of a disease would change when a new therapy is introduced, it is plausible that the way the disease is diagnosed and patients are recruited for treatment do.

An extension of a 'before-and-after' design is the use of what are known as historical controls. In this case an investigator may have a group of patients on a test therapy, and chooses a comparable group of patients with the same disease treated in the past by a different (comparator or control) treatment.

A case-series may report that a particular compression bandage in patients with venous leg ulcers has been tried and has achieved excellent results. There are many criticisms of this design. Firstly, we do not know how the patients have been selected; the clinical team may have an unerring eye for selecting those patients to be given the bandage who are likely to recover anyway. Secondly, without further evidence of the natural history of the disease, we do not know whether the patients may have recovered naturally, without intervention. Thirdly we do not know whether this type of compression bandage is better or worse than any other.

A rather stronger design is a prospective one called a *quasi-experimental* design. In this patients from one clinic (say) are given the compression bandage and patients in another clinic act as a control group and get standard therapy. The difficulty here is that again patients in the different clinics may not be comparable.

A design that is often used in Health Services Research is a *community intervention* design. This is an extension of a quasi-experimental design. For example, the cure rates for chronic ulcers are observed in two clinics. A new intervention is introduced in one clinic, and after a period of time the cure rates are again measured. An important point is that the subjects at each time point are *different*. Also the allocation of the intervention to the clinic/community is done for pragmatic reasons, such as convenience.

EPIDEMIOLOGICAL

Suppose we wish to investigate the link between chronic cough and smoking. The strongest design would be to choose a group of people, initially free of cough, some of whom were smokers and follow them up for a number of years and see how many develop a cough. This design will conform to the first Bradford-Hill criterion, in that it can test temporality. A weaker design would be a case–control study, which would identify groups of people with and without chronic cough and ask them about their smoking history. Another design would be to simply survey a group of people and ask them whether they have a chronic cough and about their smoking history. The problem with the case–control and survey designs is that they cannot properly test temporality – coughers might choose to smoke to soothe their throats! The weakest design would be a case-series whereby an investigator, say, notes that a series of people who consult about the cough appear to have a high likelihood of being smokers.

1.5 BIOLOGICAL VARIABILITY

Measurements made on human subjects rarely give exactly the same results from one occasion to the next. Even in adults our height varies a little during the course of the day. If one measures blood sugar levels of an individual on one particular day and then again the following day, under exactly the same conditions, greater variation in this than that of height would be expected. Hence were such a subject to receive an

intervention (perhaps to lower the blood sugar levels) before the next measure then any lowering observed could not necessarily be ascribed to the intervention itself. The levels of inherent variability may be very high so that, perhaps in the circumstances where a subject has an illness, the oscillations in these may disguise, at least in the early stages of treatment, the beneficial effect of the treatment given to improve the condition.

With such variability it follows that, in any comparison made in a biomedical context, differences between subjects or groups of subjects frequently occur. These differences may be due to real effects, random variation or both. It is the job of the experimenter to decide how this variation should be taken note of in the design of the ensuing study, the purpose being that once at the analysis stage, the variation can be partitioned suitably into that due to any real effect of the intervention or real difference between groups, from the random or chance component.

1.6 STATISTICAL CONSIDERATIONS

STATISTICAL MODELS

Whatever the type of study, it is usually convenient to think of the underlying structure of the design in terms of a statistical model. Once the model is specified the object of the corresponding study is then to estimate the parameters of this model as precisely as is reasonable.

Suppose in a particular experiment, we believed that an outcome y is related to the input x by means of the linear equation

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{1.1}$$

In equation (1.1), β_0 and β_1 are constants and are termed the parameters of the model. In contrast, ε represents the noise (or error) and this is assumed to be random and have a mean value of 0 across all subjects studied, and variance σ^2. The object of a study would be to estimate β_0 and β_1 in this relationship although often β_1 is the main concern. We write these estimates as b_0 and b_1 to distinguish them from the corresponding parameters.

In a laboratory experiment x might be the amount of an allergen injected under the skin and y the area of the wheal that develops. If the allergen injected results in a wheal in all subjects, but the amount injected does not influence its size, then $\beta_1 = 0$ in equation (1.1). In a clinical trial, x might take values 0 and 1 corresponding to the control and test treatments under study. In this case the null hypothesis of $\beta_1 = 0$ corresponds to no difference in efficacy between the two treatments. For an observational study y might be the diastolic blood pressure (DBP) of the individuals concerned and x their corresponding salt intake in the year before the DBP was measured. In this case, $\beta_1 = 0$ implies that the salt intake does not influence the subsequent DBP.

On the basis of this model, the two fundamental issues in an experiment to consider are:

(1) What levels of the independent variable x to choose?
(2) How many experimental units to observe?

DESIGN EFFECT

The aim of a study is to obtain as good an estimate of β_1 as possible. This implies that, for the design values $x_1, x_2, \ldots, x_N$ under experimental control, we choose their values so that the associated variance of b_1, $\mathrm{Var}(b_1)$, or equivalently its standard error, $SE(b_1)$, is as small as is reasonably possible. The variance of b_1 is expected to be

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{S}, \tag{1.2}$$

where

$$S = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

and N is the number of experimental units in the particular study. A measure of the efficiency of a particular design

$$E = 1/\mathrm{Var}(b_1). \tag{1.3}$$

Thus the smaller $\mathrm{Var}(b_1)$ the larger E and so if the values of x are under our control, we might choose them when planning the study to minimise $\mathrm{Var}(b_1)$. This choice is equivalent to choosing them in such a way as to maximise S.

In a design with values of x constrained to be within two limits (say) x_{Min} and x_{Max}, then to minimise $\mathrm{Var}(b_1)$, we would choose half the x's to have the value x_{Min} and half to have x_{Max}. This implies that

$$S = N(x_{\mathrm{Max}} - x_{\mathrm{Min}})^2/4, \tag{1.4}$$

and so

$$E = \frac{N(x_{\mathrm{Max}} - x_{\mathrm{Min}})^2}{4\sigma^2}. \tag{1.5}$$

Thus E, the efficiency, gets larger, as $(x_{\mathrm{Max}} - x_{\mathrm{Min}})$ increases.

For a given resource, one can get the most from a study by choice of a good design. The relative efficiency of two designs, I and II, addressing the same question is expressed by the ratio of their efficiencies, and is termed the design effect (DE), that is

$$DE = \frac{E_{II}}{E_I} = \frac{1/\mathrm{Var}(b_{II})}{1/\mathrm{Var}(b_I)} = \frac{\mathrm{Var}(b_I)}{\mathrm{Var}(b_{II})} \tag{1.6}$$

Suppose we were conducting a trial of a new drug at dose d, and plan to compare this with a placebo (zero dose). In this situation, $x_{\mathrm{Min}} = 0$ and $x_{\mathrm{Max}} = d$, then from equation (1.5) $E_I = Nd^2/4\sigma^2$. Alternatively we may choose a lower dose, say $d/2$, for comparison with placebo from which $E_{II} = (Nd^2/4)/4\sigma^2 = Nd^2/16\sigma^2$. Now comparing the two designs, equation (1.6) gives

$$DE = E_{II}/E_I = \frac{Nd^2}{16\sigma^2} \Big/ \frac{Nd^2}{4\sigma^2} = \frac{1}{4}.$$

This suggests that the second design is less efficient than the first, even though it is using the same number of experimental units, N.

PREDICTED VALUE

Another way to choose a design is to consider how precisely the predicted value of y at a particular value of x is estimated. That is, once the experiment is complete and we have $y = b_0 + b_1 x$ as our estimate of equation (1.1), the object is to estimate (or predict) the value of y for a given $x = x_0$ say. This gives the estimate as $y_0 = b_0 + b_1 x_0$ and this has variance

$$\mathrm{Var}(y_0) = \sigma^2 \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S} \right]. \tag{1.7}$$

It can be shown that the 'best' design, that is the one with the minimum variance, again puts half the observations at x_{Min} and half at x_{Max}. This variance is further reduced if the value of x_0 is set equal to $\bar{x}$, which in this case is midway between x_{Min} and x_{Max}. In this situation, equation (1.7) gives $\mathrm{Var}(y_0) = \sigma^2/N$. In contrast, even if the design keeps half the values at x_{Min} and x_{Max}, but x_0 is then set as either x_{Min} or x_{Max} then equation (1.7) is maximised as $\mathrm{Var}(y_0) = 2\sigma^2/N$ or twice the minimum possible value.

Amongst designs that choose different values of x, ones that set the values of x to give the minimal possible variance of an estimate are described as *optimal*.

VERIFYING THE MODEL

A crucial assumption in the above design process is that the supposed linear relationship between y and x of model (1.1) is the true one (or at least close to it). If we are uncertain about this, and this will often be the case, then it would be sensible to plan for observations in the middle of the range of x as well. Thus if we wished to try and test the linearity of the relationship a good design would be to choose equal numbers, $m = N/3$, of experimental units at x_{Min}, x_{Mean} $[= (x_{\mathrm{Max}} + x_{\mathrm{Min}})/2]$ and x_{Max}.

Example *– dose response – hepatocellular carcinoma*

In a randomised trial of the use of tamoxifen in patients with inoperable hepatocellular carcinoma, Chow, Tai, Tan *et al.* (2002) randomised patients to $x = 0$, 60 or 120 mg daily in the ratio of 2:1:2. At the design stage of the trial, it was anticipated that the highest tolerable dose of tamoxifen would bring the greatest therapeutic gain. However, there was also concern that additional activity might be slight above a threshold dose level and, should near-therapeutic benefit be demonstrated with a lower dose, this would be desirable – both in cost terms and potential side-effects. This is why the intermediate dose of 60 mg daily was added to the design. In the event, tamoxifen brought no survival advantage for these patients. Indeed there was evidence for declining survival with increasing dose.

In practice, optimal designs, such as these, are not commonly chosen except in clinical trials because experimenters have numerous, sometimes unstated, aims, and so

choose designs that try and compromise between them. Thus a design that allocates equal numbers of subjects to a wider set of x's may not be the most efficient in terms of getting the smallest Var(b) but enables the investigator to explore the responses y over a range of x values. However, in a clinical trial, the theory of optimal design suggests that if we believe in a dose–response relationship between a drug and a response, but the main concern is to show that the drug works, then one should choose a two-arm trial. This trial would compare a zero-dose control group, and an intervention group with the maximum tolerable dose.

For an observational study, it may not be possible to manipulate the x values directly, but one can often *choose* subjects who are likely to have a wide variety of these values. Thus, if we were interested in looking at the relationship between salt intake and DBP, we might choose to investigate subjects likely to have a low salt intake and compare these with subjects likely to have a high salt intake, perhaps chosen from geographical areas whose use of salt is known to differ. Within each intake group (say, low and high) there would be similar but not identical intake values.

STUDY SIZE

Although the DE may lead one to choose one design as opposed to another, it is still necessary to decide how large the study should be. This may be done by choosing the number of observations, N, to get the variance within desirable limits which have to be set by the investigating team. This implies that we may choose N to provide a specific value for the $SE(b)$ or equivalently the width of the associated 95% confidence interval (CI) for β. In Chapter 3 we describe in general terms details of how study size may be determined and for specific designs in other relevant chapters.

1.7 INFORMED CONSENT AND ETHICAL APPROVAL

It goes without saying that, before any study can take place, individual subjects have to be identified, and formal processes for their consent will have to be instituted. Clearly, the precise details will depend on the type of study contemplated, for example, whether it involves an invasive procedure, involves completing an epidemiologically based questionnaire received through the post or has therapeutic intent.

It is also usually a requirement, although again details will vary, that all studies of whatever type involving human subjects require ethical approval before they can be carried out. In certain circumstances, these considerations may have major impact on the study design. Thus a preclinical study considering the same question in man, as one that has been asked in animals, may not have the same design. For example, in a dose-finding study the dose range for man may have to avoid low doses (as they would bring no prospect of therapeutic benefit) and high doses (as they may be potentially life-threatening). The measure of drug activity is also likely to be different.

In some countries such as the UK, studies may also be subject to research governance. This means that the studies must be scientifically valid, and have mechanisms in place to ensure that they are properly carried out, written up and

Table 1.4 Major components of a clinical protocol (based on Collins, 2001; reproduced by permission of John Wiley & Sons Ltd)

	Abstract	
1.	Background	
2.	Purpose	
3.	Methods	
		Hypotheses
		Subjects
		Interventions/Comparisons
		Design
		Number of subjects
		Analysis
4.	Recruitment	
5.	Ethics	
6.	Organisation	
7.	Study forms	

disseminated. Investigators are advised to make themselves aware of the local regulations in all of these respects at the planning stage of their study.

1.8 THE STUDY PROTOCOL

For any clinical study, the main features of the study from design to analysis will have been discussed in detail at the planning stage. It is advisable to put a summary of these into a protocol which can then provide a record and reminder of the principal features of the study. Indeed Lassere and Johnson (2002) argue that a formal mechanism for making (trial) protocols, and any amendments thereof, routinely available for examination. Although details will change from study to study, there are common items for most protocols and these are listed in Table 1.4.

The Background provides an in-depth summary with references to relevant published work. Essentially this would contain the information necessary for the Introduction that will be needed for the future paper describing the study results. The purpose of the current study and its importance would be described. The Methods section should address the (major) hypotheses under test, the statistical design, the precise types and numbers of subjects who will be investigated, the interventions they will receive or the comparisons to be made and an indication of the form(s) of statistical analysis. Again, these sections should be at least detailed enough for the subsequent journal submission. This section should also include practical details of how, and from where, the potential subjects are to be identified and screened for entry and the consent procedures.

If the study is multi-centre in nature it will usually be important to describe the relevant responsibilities with details perhaps of how subjects are registered and their progress (through the study) monitored. This section may include such routine details

as contact telephone numbers and email addresses. Since recording the information is so important, inclusion of the study forms into the protocol itself is desirable, even if they are quite simple in structure. Finally the protocol should be dated, bound in book form and any subsequent amendments carefully documented. For clinical trials 'Good Clinical Practice' as described by EMEA (2002) will dictate in full the items that are mandatory for such a protocol.

1.9 REPORTING

GUIDELINES

Although we are concerned with aspects of design over a wide range of studies extending from preclinical to large-scale randomised trials and epidemiological studies, it is clear that these studies have to be analysed and interpreted and the conclusions reported. The research is not complete without this final step. Several guidelines, and associated checklists, have been published to assist authors in preparing their work for publication. These guidelines outline the essential features of such reports; in particular they clarify how aspects pertinent to their (statistical) design should be described. Just as an investigator may have a target journal in mind even in the early stages of planning a study, and thereby take note of any journal requirements concerned with aspects of their potential study, it is prudent for the investigating team to cross-check the intended design against these requirements. Anything overlooked at the design stage can then be taken account of in a design modification *before* embarking on the study. In contrast, it is too late to discover such an omission at the time of analysis and reporting.

Guidelines for reporting also give hints on what seemingly extraneous detail information needs to be collected during the experimental process. This may include the details of the consent procedures, or of outcomes in subjects who do not fully comply with the experimental process.

For those studies that do not fit into specific guidelines, it is nevertheless useful to cross-check aspects of design with available guidelines. In these circumstances, it may be useful for an investigator to compile their own checklists that can be updated by their own experience once the study is complete. Such a personal checklist will be a useful guide for the next study.

For certain types of study, including those used in the development stages of a new drug, there may be mandatory guidelines imposed by the regulatory authorities. These may set minimum standards or very specific requirements. Any investigating team ignoring such advice would need to provide cogent reasons for departure. Such departures may be entirely appropriate as new information and new situations are always arising. Should these occur then cross-checking with the regulatory bodies at the design stage is clearly prudent. For non-regulatory situations, teams may be free to have a more flexible approach. However, although flexibility is desirable, care should be taken to ensure this does not lead to lower standards.

Human studies (particularly clinical trials) have the highest standards for reporting. Thus many leading biomedical journals have adopted the CONSORT statement of Moher, Schultz and Altman (2001) which outlines the requirements for reporting clinical trials. This contrasts with publications in the experimental literature where, for

example, aspects of the choice of study design and justification for experimental unit numbers are often poorly substantiated.

STANDARDS

The second aspect of reporting is the standard of reporting, particularly the amount of necessary detail given in any study report. The most basic feature that has repeatedly been emphasised is to give numerical estimates (with confidence intervals) of comparisons made and not just *p*-values. Guidelines for referees of clinical papers have been published in several journals. These include those of the *British Medical Journal* described by Altman, Gore, Gardner and Pocock (2000). These are clearly useful for those who are designing studies, as these will eventually become the authors who are then exposed to the peer review system of the journal concerned. They would clearly benefit from knowledge of exactly what a referee will be looking for.

As indicated, the statistical guidelines referred to, and the associated checklists for statistical review of papers for international journals (Gardner, Machin, Campbell and Altman, 2000), require confidence intervals (CI) to be given for the main results. These are intended as an important prerequisite to be supplemented by the *p*-value from the associated hypothesis test. Methods for calculating CIs are provided in many standard statistical packages as well as the specialist software of Altman, Machin, Bryant and Gardner (2000, Chapter 17).

EVIDENCE-BASED MEDICINE

Following established guidelines and adopting a high standard of reporting of clinical studies of whatever type, clearly helps the reader to better appreciate the clinical messages suggested from the work that has been conducted. This in turn allows the reader to determine the relevance of the results to his or her clinical or research practice. What is more, this clarity facilitates those who are conducting systematic reviews to readily identify the key features of the study conducted for their overview, ultimately leading to more reliable synthesis and a firmer basis for EBM.

Key features

Review criteria for causality
Strength of the evidence is related to the choice of design
Check the local regulations for ethical approval and informed consent
A written study protocol
Cross-check the design with published guidelines and checklists
Ensure the reporting is to the highest of standards

1.10 TECHNICAL NOTES

Optimal Designs

Equation (1.1) can be generalised to situations in which there are more terms on the right-hand side, for example, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_v x_v + \varepsilon$. Further the form of the variable y (or a transformation of it) can be extended to binary, categorical, ordered categorical or survival time data. These correspond to logistic regression, multilogit regression, ordinal regression and Cox proportional hazards regression models. In each case the design that minimises the determinant of the covariance matrix, consisting of all the variance and covariance terms of the estimates of the parameters β_0, β_1, $\beta_2, \ldots,$ β_v, is termed D-optimal. For example, D-optimality allows for Var(b_0), Var(b_1) and Covariance (b_0, b_1) and not just Var(b_1) as we have in our exposition.