The design of empirical studies: towards a unified view

David R. Cox

Abstract A broad review is given of the general principles underlying study design with emphasis on applications in medical and epidemiological contexts. The main theme of the paper is that, while the distinction between interventionist studies, that is experiments, and purely observational ones is important, there are many common threads. A wide range of specific applications are used in outline to illustrate the discussion.

Keywords Bias · Case–control study · Clinical trial · Cohort study · Confounding · Generalizability · Longitudinal study · Metrology · Objectives of study · Prospective study · Randomization · Retrospective study · Specificity

Introduction

Many subject-matter research journals contain statistical analyses, ranging from simple graphs and tables, \( p \) values, sometimes appropriate and sometimes not, through to quite complicated model fitting, the last hopefully clarifying complex patterns of dependence. The research literature contains much less about the statistical aspects of study design. This is no doubt partly because space is short and authors are keen to set out their conclusions rather than dwell on how they got there. A more detailed reason is that the literature on study design is fragmented between at one extreme that on sample surveys and at the other that on the design of experiments, clinical trials for example, with both prospective and retrospective longitudinal observational studies in between. While the distinction between interventionist studies, that is experiments, and purely observational investigations is indeed very important the present paper aims to give a largely unified view of at least some of the main themes involved in planning an investigation.

The broad plan of the paper is that first the main phases of an investigation are outlined, the distinction between a number of modes of investigation is sketched and then discussed in more detail. Objectives of study design are summarized. The illustrative examples are intended to illustrate specific points and have deliberately both been simplified and also chosen mostly to be quite old, making omission of detail hopefully more acceptable. The paper concentrates on broad issues; there is little discussion of special procedures of, for example, experimental design or sampling technique, valuable though special methods may sometimes be.

The paper does not discuss methods for the analysis of data. An important aspect of design is the checking that incisive analysis of the resulting data will be possible, so that the study is capable of addressing the questions of concern. This usually involves formulating in advance a fairly detailed specification of the analysis proposed. This may have to be modified; for example, it may be clear that minor changes of the plan are required, such as to adapt to a small but clear non-linearity in a pattern initially expected to be linear. Much more importantly, a totally unexpected conclusion may seem to have emerged. Such situations clearly need much caution in their interpretation and if at all possible independent confirmation. It would, though, be very dangerous to ignore such aspects just because they were unanticipated. An example is outlined in the “Prespecification of analysis” section. In any major study the patterns of response that might occur should be
considered in advance and, so far as is feasible, it should be checked that information for the underlying interpretation of each potential pattern will be available.

The main emphasis in the paper is on studies of modest size with relatively limited, even if multiple, objectives. The paper does not address specifically design issues connected with so-called big data. Aspects concerned with the choice and relevance of study populations and with the quality of the data are nevertheless likely to be broadly similar to those discussed here, with quality and possible self-selection being a special concern with especially large sets of data. A separate and largely unresolved issue for big data is that of securing information that will allow the realistic assessment of precision. Simple estimates of precision based on implicit assumptions of stability and the statistical independence of individuals are likely for big data to give standard errors of estimates that are far too small. Design issues to achieve more realistic estimates of precision may include achieving homogeneous subsets of individuals. Comparison of estimates from these subsets may be a relatively realistic guide to the statistical estimation errors involved.

As in most discussions coming from a statistical base the emphasis is on the design of investigations having a viable interpretation largely or wholly on their own. Desirable or even essential though such isolated interpretations may be, in many situations major progress comes not only from considering also related similar studies but from combining information of very different kinds. For a striking example connected with smoking and lung cancer, see [1].

**An initial formulation**

A key issue is the precise formulation of the research questions to be addressed. This is particularly difficult when there is little previous work in the field. Further, some compromise is often needed between the need for simplicity and the general desirability of asking several questions together rather than in separate studies, especially when each investigation takes a considerable time to complete and analyse. Also, in practice the precise formulation will often evolve somewhat during the discussion of the design, in particular in the light of issues of implementation that may emerge only later in the development of a detailed protocol. In some extreme situations modification of the research questions during the implementation of the design may be needed, for example because of unanticipated safety issues or because important new information becomes available. Clearly any such later changes need very careful consideration if compromise to the whole study is to be avoided.

Next, how is a population of individuals for study to be defined?

**Illustration 1** In a nutritional investigation in the early 1930’s, the Lanarkshire milk experiment, the study population consisted of all children in certain classes in a number of primary schools in what was then a very deprived area of Scotland. The object was to compare children given free milk with a control group. The statistician ‘Student’, the pseudonym of W.S. Gosset, subsequently wrote a critique [2], outlined in the “Avoidance of systematic error” section below, of the implementation of the design. He also commented that it would have been wiser to confine the study to the comparison of homozygotic twins, one given milk and the other not. This would have given higher precision at relatively much lower cost. It would, however, have raised the question: how reasonable is it extend conclusions from twins to the general population? In broad terms, this was a context in which there was virtually no previous experience and the cost differences very considerable and in such situations it seems wise to regard issues of representativeness as secondary and to first study the research questions in the simplest and clearest context available.

Many considerations may influence choice of study population, accessibility, for example through being patients at a particular hospital, likelihood of showing interesting outcomes, and of well representing the target group for some health policy. Explicit specification of the study population is very desirable even if the choice may be forced to some extent by convenience. Study populations defined by self-selection are often prone to hidden dependencies distorting the conclusions. In investigations continuing over an extended time, attrition of the study population is a special concern. To the extent that the reason for drop-out can be identified some allowance by statistical analysis may be available, so that plans to identify the reasons behind potential drop-outs may be helpful.

Once focused research questions and study population have been set out, many, although not all, investigations can be encapsulated in the following scheme or by some generalization of it, very particularly to studies in several phases. Specification is required of

- units of study
- intrinsic features of those units
- exposures, in some contexts called treatments
- outcomes.

The research questions are now taken to mean: how does the exposure experienced by a unit affect its outcome, bearing in mind the intrinsic features of the unit? In most contexts the objective is to find strong stable dependencies between exposure and outcome and, so far as is feasible, to
interpret these. Sometimes, however, the emphasis may be rather more on identifying the patient or in general unit intrinsic features that suggest the most appropriate treatment for that type of individual.

A unit of study corresponds to the smallest subdivision such that two different units might, conceptually at least, have had different exposures.

Illustration 2 A unit might be a patient, a family or a community. In an ophthalmological study in which the two eyes of a patient might receive different treatments an eye of a patient would be a unit, whereas if the same patient might receive a different allocation in different periods a patient-period combination would be a unit of study. In relatively complex situations there may be a hierarchy of units, for example at one level a patient and at another a community as a focus of an educational campaign.

Intrinsic features of a unit are those to be regarded in the current context as fixed and defining the individual unit in question.

Illustration 3 In almost all contexts gender is to be regarded as intrinsic. An exception is in studies of employment discrimination where the implicit question concerns how this person would have been paid or employed had they been of a different gender, other things being equal. In this particular setting the intrinsic features are age, work experience and so on. In many clinical trial contexts the intrinsic features are those characteristics defined or measured before entry into the study. Socio-economic class or educational background of the parents would typically be treated as intrinsic. Features of the study individuals to be considered intrinsic would depend on context.

There are two broad roles for intrinsic variables. One is as a base for explaining some of the variability of outcome. In general this is achieved either by stratifying the units of study into relatively homogeneous sets within which comparisons may be made or by some method of statistical analysis that achieves that same aim indirectly. The second role is to examine the stability of exposure effects. These issues are discussed later in the paper.

An exposure of a particular unit of study is an aspect which we study for its potential impact on outcome. We consider implicitly or explicitly the question: what if the exposure had been different from how it is, other things being equal? This last qualification is taken to mean that the intrinsic features are fixed.

Illustrations 4 Exposures are typically medical treatments, or work-place or environmental hazards or life-style features, such as smoking, or aspects of diet. To be useful they need to be quite specifically defined.

Outcomes are the primary measures of features explanation of which is required. Particularly in complex studies in which units of study are followed for a substantial time period, some outcomes may primarily be of intermediate interest, for example as stepping-stones to explanations of the final outcome.

Illustrations 5 In many medical contexts survival time is a prime outcome, possibly simplified to survival for a pre-specified period, 5 years perhaps. Health-related quality of life, pain, and features of sleep are outcomes each of which may or may not be captured in a single index. Variables plotting the course of a disease, such as tumour size, may be important outcome variables at particular stages of investigation and at another later stage may be explanatory, say to survival.

The stages of an investigation may often be summarized in the following way:

- formulation of questions
- study design
- metrology
- data collection
- data analysis
- interpretation.

The present paper is largely concerned with the first three.

Types of investigation

In this section we outline and illustrate a number of key types of investigation. These are

- descriptive cross-sectional studies
- analytical cross-sectional studies
- prospective observational studies
- retrospective observational studies
- interventions, that is experiments.

The first of these commonly lacks the analytical purpose of the other types but may form an important component of a large investigation. Complex investigations may well involve a combination of types.

Descriptive study

Suppose that there is a defined population of units some feature of whom could in principle be observed for every unit. Instead of complete observation suppose that it is required to choose a sample of individuals with the object of estimating the population feature of interest.

Illustration 6 The administrator of a large hospital wishes to know the number of beds available and the number occupied at a particular reference point in time. The first step is to define the notions of bed and occupied bed and then to form a list of all available beds, a so-called frame. The target is now uniquely defined and could in principle be exactly determined by accurate observation of each bed. The planning issue is to specify the choice of a
sample of beds observation of which will lead to an estimate of occupancy with adequate precision. The most apparent reason for using sampling rather than complete enumeration is economy of effort. In addition, especially in a more complicated situation, it may well be that by greater attention to detail higher quality of data could be achieved in a sample than in large-scale enumeration, so that ultimately accuracy is higher from the sample.

There is a general implication especially for large studies, whatever their type, and for investigations involving repeated observation over a period of time: it is not always advisable to measure every feature on every study individual on every occasion. In addition to securing economy in the use of resources, it may be that taking a sample of features of subsidiary importance to the study, especially if their measurement is expensive or intrusive, is the most effective approach to secure conclusions.

There is a substantial specialized literature on sampling design.

**Analytical cross-sectional study**

In some situations observations on explanatory variables and outcomes are taken at the same point in time and refer to properties at that time. This, except in the presence of strong external information, makes for substantial ambiguity of interpretation. In particular it may be unclear which variable is explanatory and which is an outcome or whether the two variables should be regarded on an equal standing.

*Illustration 7* In a study at University of Mainz [3], patients attending a clinic for type 2 diabetes *at the same visit* had their glucose level measured and were given a series of psychometric tests assessing their knowledge of the disease and their attitude to it. In addition there were intrinsic variables, such as age, gender, educational level and the number of years since initial diagnosis. The objectives included assessment of the possible advantages of an educational programme for newly diagnosed patients. Because the glucose level and the psychometric features were recorded simultaneously it is in principle impossible from the data alone to determine whether improved knowledge leads to better glucose control, whether improved glucose control leads to greater interest in and knowledge of the disease or whether both aspects are best regarded as separate consequences of underlying socio-educational features. In some situations there may be a strong basis from previous studies or solid subject-matter knowledge for assigning a direction to any interdependence found.

**Prospective observational study**

In a prospective observational study, the units of study are individuals who have various features recorded at entry and are followed forward in time and outcome measures then recorded. Key aspects are thus the choice of individuals, units of study, for inclusion, intrinsic features of those individuals, explanatory variables both at entry and possibly subsequently, and outcome measures. Notable examples are the Framingham study, the Million Women study and the Rotterdam study. Here we outline an early and influential example.

*Illustration 8* Doll and Bradford Hill, in an investigation [4] starting in the 1950’s, took British male general practitioners as the study population and asked them to complete a relatively simple questionnaire about, in particular, their tobacco smoking. The primary outcome was survival. An account of the provisional conclusions was given about every 10 years with the final one about 50 years later. The study was observational rather than experimental because the primary exposure was observed as it was and not assigned by the investigator and prospective because the exposure was observed first and then the unit of study, a doctor, followed until an outcome or the end of the period of analysis completed. For some doctors the primary exposure, tobacco smoking, changed during the study and one of the conclusions was that cessation of smoking decreased or even eliminated the associated risks.

Investigations continuing over a long time may well involve making repeated measurements on the same study individual. Important detailed issues of design arise in choosing the time intervals between successive observation points and choosing which particular variables should be measured on each occasion.

**Retrospective observational study**

Studies of the kind outlined in the previous subsection may take a long time to bring to a conclusion. In such situations a retrospective approach may be more fruitful, at least initially. Then the starting point for each study individual is the outcome and one enquires backwards into the history of potential explanatory variables. Such an approach may lead to potential conclusions much more speedily than a prospective study. A main limitation of the retrospective approach is the general frailty of information collected about the past, especially the remote past, enhanced by the possibility of selective bias from recall, the outcome being known.

A further consideration is that prospective studies concerned with a rare outcome are likely to be inefficient in that an unnecessarily large amount of data is collected on individuals not experiencing the critical outcome of concern. To some extent this may be obviated by having multiple objectives. A typically much quicker route is to start with the rare cases, choose controls, possibly matched case by case, and enquire into the history of potential
exposures. This is the so-called case–control method. The challenges with the approach may stem from the difficulty of specifying an appropriate control group, but more particularly, as noted above, from the possibility of selective recall when information about earlier exposures is collected.

Illustration 9 Doll and Hill [5] took as cases all patients with carcinoma of the lung at 20 London hospitals over an 18 month period. For each such patient there was a control from the same hospital, age-band and gender. All patients were interviewed by a hospital social worker about their smoking and about other aspects. After some exclusions there were about 700 patients in each group. The paper contains searching discussion of the possible biases in such a study.

Experimental study

The studies described above are observational. Even though the investigator may have complete control over which units of study are used and which features are measured, the central aspect, namely as to which exposure is experienced by a particular study unit, has to be taken as prespecified at its observed level. By contrast, in an experiment the investigator determines by an explicitly formulated algorithm which exposure is to be used for each particular unit. The primary distinction is not between the use or not of randomization to determine the allocation but between the absence or presence of specification of the choice of allocation by the investigator. In the great majority of situations an element of randomization is very desirable, or even close to essential, but there are a few situations where randomization is a bad idea: we discuss these in the “Randomization” section.

Illustration 10 In an early investigation of the use of beta blockers in secondary prevention of heart attacks, patients, having given informed consent, were randomized between receiving a beta blocker and receiving medication superficially identical, but in fact a placebo. The objective was that neither the patient nor the treating physician was to be aware of the treatment allocated to each specific patient; here the exposure under discussion is an allocated treatment. At a later stage some test results from each patient were sent in randomized order for subjective assessment by an independent expert. The primary outcome was weight gain over a several month period from winter to summer. ‘Student’ [1] showed that the weight gain of the better off children was relatively depressed and that the children from the relatively more affluent families had lighter clothing in summer. Thus the weight gain was virtually uninterpretable. Weights were measured with the same weighing appliance. The poorest children had only one set of clothes, the same in winter as in summer, whereas the children from the relatively more affluent families had lighter clothing in summer. Thus the weight gain of the better off children was relatively depressed and these children had been selectively assigned not to receive milk.

Generalizations

In the previous subsections a rather stark outline has been given of a number of types of study, all except the first aimed at a similar objective of studying the dependence of an outcome on explanatory features, called here exposures. Sometimes, especially in studies continuing over a substantial time period, mixtures of the primary types of investigation may be used, for example a case–control study embedded in a prospective cohort or within a randomized experiment in order to explore an unanticipated effect that becomes apparent only at the intermediate stages of the primary study.

Requirements

The primary requirements of a study, once the objectives and study population are specified, are that the conclusions should be free of systematic distortion and that random errors of estimation of effects defining the conclusions should be controlled to a reasonable level. It may often be more effective to study several questions in one investigation than to proceed in entirely separate steps. This may be called broadly the factorial principle and is discussed in the “Factorial principle” section.

Avoidance of systematic error

Illustration 1 ctd In the Lanarkshire milk experiment outlined above there were essentially two possibilities for each child taking part, free milk each school day and a control, no free milk. The children were divided into two groups by the head teacher, either by lots, that is at random, or in alternation of alphabetical order of their surname. Then the head teacher was allowed to adjust the two groups to correct for any imbalance perceived, that is to “improve” the randomization. The primary outcome was weight gain over a several month period from winter to summer. ‘Student’ [1] showed that the adjustment to the randomization had resulted in a substantial initial imbalance in the two groups, the especially impoverished children being more likely to receive milk. There was a further aspect that made the study virtually uninterpretable. Weights were measured with the same weighing appliance. The poorest children had only one set of clothes, the same in winter as in summer, whereas the children from the relatively more affluent families had lighter clothing in summer. Thus the weight gain of the better off children was relatively depressed and these children had been selectively assigned not to receive milk.
The role of randomization in study design is discussed in more detail in the “Randomization” section. In the illustration just given its role is specific to experiments as contrasted with observational investigations. Another main technique for avoiding bias is that of comparing like with like, either directly by matching pairs or groups of individuals, as suggested by 'Student' in the Lanarkshire milk experiment, or indirectly by statistical analysis, provided appropriate data are available. A more specialized method exploits the special independency structures involved with instrumental variables, often in the form of Mendelian randomization.

Illustration 1 concerned systematic errors arising from defective technique. In many situations, especially for observational studies, there may be particular concern with the influence of unobserved variables. There are broadly two aspects. An apparent effect of an exposure may really be the consequence of an unobserved feature of equal status, that is the two variables refer to the same time point so that neither is explanatory to the other.

Illustration 11 In early work on the incidence of HIV-AIDS, the number of sexual partners and the use of amyl nitrite pills were possible explanatory features, strongly associated with one another. Had the former not been available the false conclusion would have been drawn that the outcome was influenced by the use of amyl nitrite; number of partners would have been an unobserved confounder. The discussion was complicated by a more subtle issue [7]. The former explanatory variable was measured much less reliably than the latter. This has the consequence that unless appropriate analysis is used dependence is apparently shifted from the poorly measured explanatory variable to the strongly associated well measured but in fact irrelevant correlate.

A second possibility, called confounding in the observational literature, is that the apparent effect of a particular explanatory variable on outcome is explained by an observed or unobserved variable explanatory to both the specific observed explanatory variable and to the outcome, in extreme cases accounting for the whole of the effect. If the particular variable is observed careful statistical analysis may largely remove the distortion, but the possibility of serious distortion from unobserved features is a common concern with observational studies, especially when relatively modest effects are under study. This concern is avoided by randomization, at least as regards features defined before randomization, even if unknown. More precisely, the possibility of such an effect occurring is moved from being a systematic error to being a random error, the possibility of which is assessed in careful statistical analysis.

While it is possible that a large apparent effect of an exposure is in fact totally the consequence of other unobserved variables, this is decreasingly likely if the effect is stable across a range of intrinsic variables and is in any case less and less likely the larger the effect involved. A general question that may be asked, particularly in observational contexts, concerns what properties a confounding variable would have to enjoy to explain all or much of the dependence observed or, more alarmingly still, reverse the direction of an effect.

**Control of non-systematic error**

Haphazard or largely random variability is the focus of many statistical techniques of analysis. From a design perspective its effect can be controlled by one of three broad methods, namely improved measurement technique, internal replication and external, or genuine, replication. If the effective error arises from a number of distinct sources a combination of the three approaches is likely to be needed.

Illustration 12 Suppose that in a study, observational or experimental, each patient is X-rayed. For further analysis an expert is to attach a score to each X-ray assessing some aspect of concern. Improved technique would involve achieving greater clarity of the images. Internal replication would involve having each X-ray scored twice and external replication would involve having two X-rays for each patient. Internal replication in which the same expert saw each image twice a short time apart, knowing the second time the score from the first assessment, would be likely to produce a spurious impression of precision. If the second assessment were blind to the first assessment, or better still done by a different assessor, better estimates of reliability would be possible. If, no doubt unrealistically in this particular instance, fully external assessment were possible, two separate X-rays would be taken for each patient and assessed independently. It would be important in similar more realistic contexts that any such external replication was sufficiently close in time that no real change had taken place.

Issues such as those of the previous Illustration have to be taken in context. In many situations it may be known from previous experience that the measurement technique is secure. In others, occasional checks on a sampling basis may be sufficient. Unnecessary replication may be both immediately wasteful and a harmful distraction. Appreciable measurement error that is not accounted for in analysis may, however, seriously distort the conclusions. In simple situations the apparent relation is attenuated, that is too weak, but, in some ways more seriously, as noted in Illustration 11, in more complex contexts a dependence on a badly measured explanatory variable may be transferred to a highly correlated, well-measured but in fact irrelevant alternative explanatory variable. If measurement error is likely to be important realistic estimation of its magnitude
as part of study design will be required. In some contexts short-term variation in an explanatory variable, such as blood pressure, may be of direct subject-matter interest not merely an indication of error.

The second aspect determining precision of the conclusions concerns the scale of effort suitable, in particular the number of independent units of study, for example patients, that should take part. Such assessment may take a number of forms. It might be known approximately how many patients are likely to be available: is this enough to give adequate precision to the conclusions? More commonly, a prior assessment of an appropriate number is required. This is often done by the statistical notion of power but an essentially equivalent, simpler and in general more useful method is to consider the precision of estimated effects likely to be achieved.

Illustration 12 In a study to assess the effect of an intervention in reducing incidence of a disease the calculation of the size of the trial aimed to make the 0.95 level confidence limits on the percentage reduction to be plus and minus 10%; that is, if the estimate reduction were to be 20% the confidence limits would be from 10 to 30%. This was close to the level of precision actually achieved. Specification of the power of the significance test of a null hypothesis would have required specification of three numbers, two error probabilities and a critical difference, many different sets leading to the same proposed sample size; moreover, while power may be relevant before the data are obtained, it is almost irrelevant once the data are available. By contrast the width of the confidence intervals achieved is directly comparable with the target value and any serious discrepancy can hopefully be interpreted.

In principle, decisions about the scale of effort appropriate involve a balance between the cost of an investigation, sometimes roughly proportional to its size, and the penalties associated with errors of estimation. In the present context, however, it is probably very rare that this choice can be usefully formalized quantitatively.

Usually the statistical error in estimating a particular effect, as measured by its standard error, is inversely proportional to the square root of the number of independent units of study involved. Large extrapolation from a small study to assess the precision from a large study may, however, be unwise. For various reasons the underlying level of haphazard variability in a large study may be greater than in a smaller study where tighter control may be achievable.

**Factorial principle**

A final general aspect likely to arise especially in larger scale studies is that it is often more effective to study several related issues together rather than by separate investigation of each. In the context of experimental design this leads to the notion of factorial experiments but the considerations involved are much broader. There are two different advantages, the possibility of deeper study of the relations between different effects and issues of economy of effort.

Illustration 13 In an investigation [8] of two treatments for pleural infection, t-PA and DNase, patients were randomized between four groups, namely double placebo; t-PA plus placebo; placebo plus DNase; both t-PA and DNase. This forms a so-called $2 \times 2$ factorial design. The conclusion was essentially that the single treatments had little effect but that there was a substantial benefit from the combined treatment. This illustrates the two different points. First, separate experiments on t-PA and DNase comparing them with a placebo would have failed to find the primary conclusion. Secondly three separate experiments each comparing one of the three active treatments with a placebo would have had two disadvantages. One is lower precision because of the distribution of effort across possibilities. Moreover the comparison of the composite treatment with its components would have been an across-experiments comparison and that might have lowered security. In statistical language the factorial experiment allows the exploration of interaction between factors. It is, however, important that finding an interaction is not a conclusion but a warning that more careful interpretation is likely to be needed than one based on the effects of the factors separately.

**Some general aspects**

**Choice of features**

We now consider some broader aspects of design, starting with the choice of features for measurement. As discussed above, variables may be classified as intrinsic variables or as exposures or as outcomes. The first two types are in a general sense explanatory to the outcome. Particularly in an investigation over a period of time variables may change their status but at any particular point of analysis, intrinsic variables are regarded as essentially fixed properties of the units of study. In investigating the effect of an exposure we are implicitly or explicitly considering the question: how might the outcome have changed had the exposure been different, the intrinsic features remaining fixed.

Illustration 14 To use age as the explanatory variable for a study of why some people’s hair turns from dark brown to grey adds nothing to the description that such changes do occur over time. A particular physiological process or a particular stress could be regarded as potentially explanatory because, conceptually at least, the process might have
been inhibited, all else remaining the same and the consequence for hair colour assessed.

Exposures and outcomes are at the heart of the research questions under study. There are several reasons for including in addition intrinsic variables. These are

- to control for unwanted sources of variation and in particular rule out alternative explanations of any strong dependencies uncovered, occasionally by assuming the intrinsic variables to be instrumental variables or more commonly by some type of regression adjustment
- to check for the stability of any exposure effects detected
- to enhance generalizability and specificity of the conclusions.

The third aspects, which are interrelated, are discussed in more detail in “Generalizability and specificity” section.

**Time scale**

An important general feature affecting many aspects of study design is the typical length of time between the initial planning of a study and its completion. When this is quite short, as in some kinds of laboratory work, the security of the conclusions from individual studies is less critical, although even here a serious design error, leading, for example, to a false warning of a concern about patient safety, may block a promising line of enquiry. When there are a large number of potentially important explanatory variables specialized techniques are available facilitating investigation in a series of steps. When, as is quite common in epidemiological investigations, a long time scale is involved, a number of different aspects are involved.

Early audit of the data collection and management aspects is important, especially if there is not a tradition of similar investigations available for guidance. Interim analysis also will be important. There are critical general issues involved to limit the frequency of such interim analyses although, provided appropriate care is taken, there is no specifically statistical reason for imposing such a limit, but modifications to the analysis, usually minor, may be required. Precautions are required to rule out the selective reporting of results. The limitations on the frequency of intermediate analyses arise more from the effort involved, and the temptation to make frequent unsettling changes in procedure based on what is found. Also in some contexts the security of an investigation would be fatally compromised if partial information about the outcomes were to become widely available. In studies with very long duration interim reports on the conclusions may be called for. Changes in study protocol may be needed.

*Illustration 15* In a study planned to recruit patients over, say, a 2-year period it might be found after a while that recruitment of patients was much slower than had been anticipated and that it was unlikely that reasonable precision would be achieved at the end of the 2 years. It might then be reasonable, after a specified date, to broaden the eligibility criteria, for example by extending the age range. It would be desirable in the analysis to check that either the patients admitted only under the extended criteria were not having an anomalous effect on the conclusions or, that if they were, that should be reported and if possible explained.

Another issue affecting studies with a long duration is that information may appear from other investigations not available at the planning stage. In extreme cases this might suggest including new variables or stopping the collection of some current variables. It is a subsidiary argument for having multiple objectives to a study that some objectives might be abandoned without the whole being fatally compromised.

*Illustration 16* In early studies of the possible carcinogenic effects of use of mobile phones the long latency period between initiation and occurrence of symptoms meant that direct investigation of, say, brain cancer incidence rates was unlikely to be effective. Instead some weight was placed on physics showing the unlikelihood of the possible penetration through the skull of effects with the wave-length of mobile phone signals.

**Metrology**

Suppose that it is required to measure a particular feature of a study individual, for example health-related quality of life. Metrology covers the issues involved in choosing a method of measurement that ideally is relatively simple to deploy and sufficiently reliable for the purpose in hand. In many contexts the development of relatively easily deployed methods of measurement is central to progress. For example, the development of chromatography for separating complex mixtures of components had a major impact on biochemistry in the 1940–1950’s and the crucial role of scanning devices in studying the brain is clear. In some fields, for example when safety issues are involved, it may be feasible to set up a special study solely to examine the sources of error connected with a particular technique although this is, perhaps, unlikely to be possible in a medical or epidemiological setting.

Nevertheless it may be possible to imbed such a study within a larger investigation, the objective being to establish the components of variability involved in a realistic setting. Thus any duplicate measurements should be made with appropriate concealment, achieved for example by randomization. In other contexts, there may be important
issues concerning the number of different components needed to capture the feature in question.

Illustration 17 Health-related quality of life may require an extensive questionnaire for its assessment, the assessment for each individual being in the form of several scores corresponding to specific aspects; for a general discussion, see [9]. In particular contexts a much simpler approach may be appropriate and indeed much more effective. Thus in a possibly apocryphal study one single question was adequate as a basis for comparing two treatments for arthritic patients: could you today put your socks and shoes on unaided? This is not intrusive and the answers are likely to be largely accurate. Measurement of sleep and of pain raises similar issues about the complexity of definition relevant in different contexts.

Role of previous experience

Entering a largely unexplored field clearly raises different issues from that of planning a new study in an already well studied area. In the latter not only the research findings but also the techniques of design, data collection and analysis may be fairly well established. Yet the dead hand of precedent can be inhibiting. It is not unknown that the first investigators in a field may use methods of study design and analysis that are adequate for their immediate task but which may be far from ideal for general use. Particularly if the pioneering study is judged important and successful, it may be difficult to gain acceptance for new and better methods.

Prespecification of analysis

A critical aspect of planning is to prepare a scheme of analysis sufficiently detailed to check that the key objectives of the investigation can be addressed. In particular the main patterns of outcome that may arise should be capable of secure interpretation. The plan may, however, have to be abandoned. A relatively minor example would arise if the initial plan involved fitting a straight line to a relation between two variables. If the data in fact show a nonlinear relation there are two broad possibilities. One is that a relatively minor adjustment to the analysis is needed without any change of broad objective. The other is that the nonlinearity is of intrinsic interest, that is has some subject-matter interpretation that needs explanation and discussion.

The more extreme possibility is that some qualitatively different and totally unanticipated outcome arises.

Illustration 18 In a conversation between two cardiologists and a statistician the two cardiologists agreed that a modification \( T \) of a standard regimen \( C \) was likely in a proposed randomized trial to show a reduction of systolic blood pressure of roughly 20–30 mmHg. When the statistician asked: what if blood pressure actually shows an increase with \( T \) not a decrease? cardiologist 1 said that in 5 min there would be an explanation and cardiologist 2 added that in 10 min they would be discussing the alternative merits of two different explanations. In fact the expected decrease occurred. The primary point is that had an increase occurred, explanations would soon have been available but would have needed independent confirmation.

Illustration 19 A study of the interaction between cattle and wildlife, namely badgers, in the propagation of tuberculosis, *Mycobacterium bovis*, compared three randomized treatments, namely a control, a reactive culling policy and a proactive culling policy [10]. Each study unit was an approximately circular area containing about 10^3 cattle and there were 10 sets of three such areas, the three areas forming each triplet being similar and quite close geographically. The different triplets were spread over the high-risk areas. In the reactive areas once a TB case was detected on a previously free farm, culling of badgers took place in a local area around that farm. In the proactive area every year for five years a team of field workers aimed to trap and shoot as many of the badgers in the whole area as possible. The outcome variable was the number of previously TB-free farms in which *M. bovis* was detected. The size of the trial was such that the standard error of the estimated reduction was expected to be about 5%. It was thought that the reduction from proactive culling might be 20–30% with that from reactive culling appreciably less. The intermediate analyses, seen only by the statisticians involved because of the high sensitivity of the issues, showed that while the proactive culling had about the expected effect the reactive policy appeared to *increase* the TB rate, a totally unanticipated outcome. That arm of the trial was converted to a control only. But the totally unanticipated outcome had to be explained. Discussion with animal ecologists and others led to two possible explanations. One was that it was a bias in that some farmers in the control areas were illegally killing badgers and doing so more effectively than was possible in the reactive strategy. The other was the phenomenon which animal ecologists call perturbation. The culling distorts the stable behaviour pattern of the badgers, they travel much further and hence propagate *M. bovis* more widely. A supplementary small scale study involving depositing different colours of peanuts outside the setts in which badgers live and noting where they are deposited showed clear inconsistency with the first explanation and consistency with the second [11].

The point of this Illustration is that unexpected conclusions may be of critical importance for understanding. Nevertheless caution is needed in two senses. There should be as much assurance as is feasible that the conclusions are not spurious, by-products of the play of chance. Also, and importantly, explanations proposed should be subject, as
Intermediate analysis

In some studies intended to continue over a substantial period there may be interim formal reports planned at regular intervals.

Illustration 8 ctd  The GP’s study [4] of smoking outlined above issued a formal report every 10 years for a 50 year period. Important intermediate conclusions emerged over the period.

More commonly, at least in smaller scale investigations, conclusions will be reported sometime after the end of data collection. Intermediate analyses, possibly relatively informal, will, however, typically be very desirable for a number of reasons:

- to check, especially in studies in a relatively novel field, that the administrative and measurement procedures involved are working correctly. This may involve a detailed audit of a small sample of individuals
- to check that it is likely that reasonable precision will be achieved by the study as planned or that such a level of precision has already been achieved, suggesting that the study should either be regarded as complete or be modified
- especially importantly, to check that there are no anomalous findings, for example unexpected safety issues.

There is no formal statistical reason why intermediate monitoring should not be virtually continuous. There may, however, be very pressing practical reasons why such monitoring is best limited. Too many modifications of the protocol may, even if individually sensible, be disruptive rather than helpful. Often more seriously, intermediate information may, misleadingly, appear to suggest, say, the superiority of one treatment regime, and if this information is widely reported, misreported or rumoured continuation of the study may well be compromised.

Sometimes what are best thought of as intermediate variables are treated as substitute or surrogate outcome variables. That is, the outcome variable of primary interest is not available and it is assumed on the basis of subject-matter knowledge that the effect of explanatory variables on the surrogate is qualitatively the same as it would be on the primary outcome.

Generalizability and specificity

Sometimes studies are designed with a very particular objective in mind, for example to settle a public health issue in a particular community. Even then, though, there may be difficulties in extending the conclusions from the data to the objective. Much more broadly there is some tension between the objective emphasized in much statistical discussion of designing and analyzing studies that in a sense stand on their own, as contrasted with the need to produce some synthesis of what is known in the field in question and of the relation between the conclusions from different studies and types of investigation. So while the emphasis in these notes is broadly on investigations that to an appreciable extent do indeed stand on their own, the issue of extending the conclusions and relating them to other work in the field is very likely to arise.

We deal with this under two superficially contrasting but in fact strongly connected themes, generalization and specificity. Generalization of conclusions, for example to a new study population, may be based on

- similarity of conclusions from related studies
- synthesis of information of different kinds
- checks of the stability of the conclusions from the study under analysis, for example that the primary exposure effects are essentially the same in two ethnic groups, are largely independent of age and socio-economic class, and so on
- explanations based on connections with an underlying biological or physical process
- generalization based on some formal statistical notion of random sampling from an underlying target population.

Illustration 1 ctd  A striking example of the importance of considering information of different kinds is the review [1], already mentioned in the Introduction, of the connection between smoking and lung cancer. Several leading statisticians of the time, from quite different perspectives, were unconvinced that there was a genuine dependence. Cornfield et al showed that when evidence of several different kinds was considered together an overwhelming case for a direct dependence emerged.

The last basis for generalization, the formal statistical one, is probably rather rarely applicable. For instance in a well-organized clinical trial the patients enrolled are those attending participating centres and willing to give informed consent. Even with perfect compliance with the treatment schedule it is unlikely that such patients are essentially equivalent to a random sample from the target population, making somewhat tenuous the connection with the formal statistical notion of generalization from a random sample. Explanations from understanding an underlying process are clearly highly desirable and may often be the most secure route to understanding of the strengths and weaknesses of conclusions. Nevertheless any such explanations produced ad hoc should, as emphasized above, be regarded as temporary and in need of independent confirmation.
Specificity is at first sight the complement of generalizability in that it is concerned with consequences at an individual level. In fact it is broadly parallel in that it involves extending an aggregate conclusion for the whole study population, in one case to a new population and in the other to a specific individual.

Illustration 21 A treating clinician has a particular patient to assess. There are two treatments available, A and B, and a recent very well designed trial has reported a clear difference with A having significantly better success rate than B. The clinician has much information about the patient; what should be done? Now the clear trial result certainly does not show that A is better than B for every patient. Its formal statement is that in two hypothetical situations, one in which all patients receive A and one in which all patients receive B, all else remaining the same, then the aggregate outcome is better in the former than in the latter. This is, of course, much less than showing that A is preferable for every patient in the study. To the extent that the conditions for generalizability hold it becomes more secure for the clinician to recommend A. In particular, the more the preference for A has been shown stable with respect to intrinsic features of the patients, the stronger the evidence that A is likely to be right for any specific individual.

Randomization

Randomization is the use of an objective impersonal process with simple agreed probability properties. It is an important technique both in the selection of samples from a specific population and in treatment, or exposure, allocation in an experimental context. Indeed the distinction drawn here between observational and experimental or interventionist investigations is often, slightly misleadingly, described rather as between observational studies and randomized trials.

The following discussion centres on randomization in an experimental context. There are broadly four arguments for such randomization. Two of these are technical statistical issues concerned with justifying statistical analyses directly from the randomization used, with minimal assumptions about the form of the haphazard variation present, as contrasted with specific assumptions about that variation. These arguments will not be considered here.

The two more empirically based arguments for randomization are that it converts what might have been systematic distortions of the conclusions into random perturbations reduced by replication and whose magnitude can be assessed by analysis. Further, if the randomization is in effect a public procedure accusations of implicit or explicit bias by the investigator are refuted. The second point is that randomization combined with appropriate concealment means that certain kinds of subjective bias are excluded.

Illustration 19 ctd The large-scale randomized trial connected with the relation between badgers and bovine TB discussed above was highly controversial with, broadly speaking, farmers on the one hand and animal rights activists on the other convinced that badgers for one and cattle for the other were the prime source of infection. As each set of three areas was randomized between the three treatments the randomization was independently witnessed to give public assurance of the lack of bias and to refute any suggestions of conscious or unconscious bias.

Randomization is in principle needed at every stage of an investigation at which appreciable error might be introduced.

There are two situations in which randomization is a bad idea. One is where a complex sequence of operations is to be carried out on each unit of material. If this sequence is constantly changing the possibility of confusion and undetected errors of assignment may outweigh any advantages of randomization. Occasional changes in sequence may provide better security.

The second situation is where very small numbers of units are involved. As an extreme example if there were just four units and two treatments or exposures to be applied in time sequence some of the arrangements would be unwise, for example putting one treatment first and the other last.

Illustration 21 It is reported that R.A. Fisher, who introduced formal randomization into experimental design, was asked at some point: what should be done if when planning a study in which a number of individuals are randomized at essentially the same time randomization throws up a pattern with special undesirable features? Fisher is said to have reacted with surprise: throw the arrangement away and start again, he advised.

It is important that this is both the common-sense and the theoretically correct solution. A realization that an appreciable number of the assignments produced by the randomization are likely to be unacceptable would signal a need to review the appropriateness of the design being used.

Conclusion

The emphasis in these notes has been on the design of an individual study to address one or more specific issues, although as noted in the “Introduction” section many of the ideas may be adapted to a larger scale investigation.

To implement effectively some of the ideas outlined in the paper may sometimes require special statistical techniques, particularly so when, for example, a possibly
complex sequence of exposures or treatments is involved over a period of time. Such more detailed aspects of the planning of investigations have not been discussed in the present paper. The references that follow include an introduction to some of the more specialized literature. For sampling, see [12] and for the design of experiments, [13–15]. For prospective studies, see [16] and for retrospective studies [17, 18]. For Mendelian randomization, see [19].

Acknowledgments I am grateful to David Allison, John Goldthorpe; Christiana Kartsonaki, Ruth Keogh, Nancy Reid, Tom Sniijders, Bianca de Stavola, Jan Vandenbroucke and Nanny Wermuth for very helpful comments. The paper is based on lectures given at London School of Hygiene and Tropical Medicine and at Erasmus MC, Rotterdam. I thank anonymous questioners at these lectures. Finally the paper benefited considerably from very constructive suggestions by the referees.

References