

Review article

Statistical issues – significantly important in medical research

M. Gellerstedt

Department for Studies of Work, Economics and Health, University of Trollhättan, Uddevalla

Martin Gellerstedt, PhD Statistician
Box 795
SE-451 26 Uddevalla
Sweden

Accepted for publication 17 September 2001

Medical researchers may regard a study mostly as a project with medical methods and issues to be dealt with by medical professionals. Nevertheless, this view may allow input from other fields when necessary. For instance, a study often includes some statistical calculations, and if an easily handled program package is not available, there may be a need for a statistician. The belief that statistics is equal to analysis leads unfortunately to the phenomenon that statisticians are contacted at the end of the study period, when it is time for 'mathematical acrobatics'. Unfortunately this is, in many cases, too late – not even the most sophisticated analyses could rescue errors made during the earlier vital steps, e.g. in the design and sample size calculation. As a consequence many publications in medical journals are statistically poor or even wrong (1–3), and many studies are not published due to statistical weakness. In short, medical researchers are often not aware of the importance and broad use of statistics in a study, even though it is clear that statistics has an important role in medical research. In 1983 it was shown that 70% of published articles in the New England Journal of Medicine used statistical analysis (4). Since then the use of statistics has increased and is nowadays more or less standard.

In contrast to this view, a statistician may regard a clinical study more as a study based on statistical concepts with clinical input, rather than a clinical study with statistical input. From this point of view it can be rather frustrating not to be part of the vital planning and design questions which, in the opinion of a statistician – to a high degree – are considered as statistical matters. Statisticians can in many cases easily find statistical errors in studies, and sometimes are even annoyed by (or find enjoyable) the lack of statistical precision.

Why these views differ and why there is a lack of cooperation is not discussed in this article. As a statistician I can only admit that statisticians have a great responsibility, especially with regard to marketing and presentation of theory in its entirety and in an understandable way. It is time to increase cooperation and to create a mutual point of view. Both medical researchers and statisticians must understand that each single step of a study needs both medical and statistical input to achieve high scientific quality. In this article some of the basic statistical concepts of a study are explained and justified. Design issues are discussed briefly while statistical significance is discussed in detail. The focus is on experimental confirmatory clinical trials on humans, but there are some comments regarding observational studies, e.g. case-control studies. The objective is to illustrate that statistical issues are important to all stages of a study. Furthermore, the intention is to give the reader an increased understanding of statistics, an interest in basic concepts, some good advice, and references to help produce high quality studies.

The importance of a good design

It is commonly stated that a good trial is a trial with an important and interesting question, which is answered with precision. In a confirmatory trial the researcher wants to use results found in a sample of patients to make generalizations to all patients. We could say it is a wish to change tense, from past tense to present tense. We want to use the results from how it *was* (for the studied patients), to stating how it *is* (generally for all patients). To be able to make that generalization – the switch of tense – it is necessary to have an appropriate

design. The justification of the analysis depends on how the data is collected (5). Unfortunately this is often a neglected part of medical research. Perhaps it is due to a pedagogical failure, where there is not enough emphasis on design issues in medical statistics' classes (6). Designing a study is not an easy task: 'There are only a handful of ways to do a study properly but a thousand ways to do it wrong' (7), and it demands a lot of effort. But if we bear in mind that it takes just one bias to make a conclusion fall flat, it is worth all the effort. I will now discuss some of the basic aspects.

Primary objective

To start with, it is extremely important to clarify the primary objective in detail. This could be done by stating a clear detailed hypothesis or by describing the effect that is going to be estimated. According to the International Conference on Harmonization (ICH) (8) 'a confirmatory trial is a trial in which a hypothesis is stated in advance and evaluated'. Of course this is important for practical reasons, but it is also important for statistical reasons. If the primary objective and a hypothesis are predefined, then the results of the analyses will have a much stronger scientific proof value. This is discussed further under 'Other dilemmas regarding statistical significance'. Formalizing the primary objective also includes defining a primary variable which should be capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial [*ibid.*]. When discussing the primary variable (efficacy variable, target variable, primary endpoint), we must consider exactly what is going to be measured and how it is going to be measured.

Assume that we want to study the lowering effect on blood pressure of a new treatment. Should we use the diastolic blood pressure (DBP) after treatment, or use the change in DBP compared to baseline? Measured sitting or lying? Which arm? How long is the treatment? Digital equipment or manual equipment? Repeated measurements on the same individual? and if so, how many? Even such a common variable as blood pressure raises a lot of questions. A statistician, used to evaluation of the quality of a variable, i.e. its validity and reliability (9), can at this stage be a valuable discussion partner. In many studies surrogate endpoints are used (10, 11). For instance, high blood pressure is not a problem in itself but it can cause many problems. Therefore we are measuring blood pressure as a surrogate endpoint for the causes, just because it is simpler and faster and does not demand as many patients as if we used cardiac infarction or mortality as the primary variables.

The reliability of the variable and the expected effects of the treatment are also important basic points for

designing the trial, namely dimensioning, i.e. calculating sample size. In an observational study, especially a survey based upon questionnaires, we have the same type of operational problems. We need to use a standardized questionnaire (12) or develop new valid, reliable questions and scales (13). In summary, to define the primary objective and formulate it with a hypothesis to test an effect, and to estimate and define a corresponding primary variable, are important issues from the beginning in a good trial. Statistical issues in this phase of the trial are, firstly, to validate the variable(s), and, secondly, to foresee possible analyses and estimate the sample size.

The use of a control group

Many researchers think of effect as the difference between the baseline value and the corresponding value after treatment. But this is actually only a special case in a situation where we can guarantee that nothing else except the treatment could have affected the variable studied. Instead, effect should be thought of as the difference between the value at baseline and the value after treatment, compared to the difference we would have seen without the treatment (14). Of course it is not possible to treat a patient and to not treat a patient at the same time. The role of the control group is to give us a chance to estimate what would happen without treatment. The true treatment effect is then estimated by comparing treatment group with control group. Thus the justification for using a control group is basically that changes in values could occur without any treatment. Changes could be due to the phenomenon *regression to the mean* (15–17). Generally there is a tendency to start treatment when the studied value is on a peak. Furthermore, it is rather common to have a cut-off limit for in studies. For instance, to be enrolled in a blood pressure study the patient may require at diastolic blood pressure of at least 90 mmHg. Blood pressure is a random variable, it rises and falls with a natural variation (18). If the patient's average DBP is 85 mmHg it is not unlikely that the pressure will occasionally exceed 90 mmHg. It is possible for this patient to enroll in a study during such a peak. A decrease in DBP after treatment could be partly or totally explained by the fact that the DBP has returned to its average level (or even a trough value). Consider a large company that suddenly notices the number of employees absent due to illness is extremely high. At that moment the company starts a campaign with a bunch of health strategies. Six months later the number of people absent due to illness is decreased to a normal level again. Was this due to those strategies or simply because of regress to the average level? There are some possibilities to analyze whether a change is due to a high initial value or not (19, 20).

Changes in values could also be due to a time trend for a specific illness. Consider a number of patients with chronic obstructive lung disease. Assume that we are studying this group for 10 years. We expect a progressive worsening of lung capacity. If lung capacity is unchanged after 10 years of treatment, the difference between baseline value and value after treatment is zero. However it would be wrong to conclude that there was no effect because we would expect a decrease in lung capacity if the patients had been left without treatment. The value under study might also be affected by placebo effects, i.e. subconscious effects related to the patients' or observers' expectations. This issue is discussed further in the section entitled 'Blinding'.

In an experimental study a control group is easy to find, for instance by randomizing half of the patients to the control group. However in some designs the choice of a control group is not always obvious. In a case-control study the choice of the controls is a crucial step which needs statistical considerations (see under 'Observational studies – case-control design'). There are other choices, such as historical controls (21, 22), which must also be evaluated to judge the possible influence of bias.

In summary, changes could arise for several different reasons; the use of a control group allows us to differentiate between changes due to treatment and changes due to other factors.

Randomization

There are three important reasons for randomization. Firstly, randomization is the most objective way to allocate patients to the active group (new treatment) or to the control group (placebo or competitive treatment). If the investigators control the allocation process there could be a selection bias, i.e. a tendency to allocate patients in a way that advantages the active treatment (23, 24). Secondly, randomization is fair in the sense that in most cases it creates groups that are equal *on average*, regarding both known and unknown factors, i.e. comparable groups. Statistical inference does not need exactly equal groups, it takes account to variability both within the groups and between the groups. As a matter of fact, most of the inferential statistics, such as tests and confidence intervals, is valid only if randomization has been used. This is the third important reason for using randomization. Nonrandomized studies demand special statistical techniques (25). Simple randomization means that each individual is independently randomized to one of the groups, most often with a uniform probability distribution to the groups. In situations where it is important to keep the groups to a similar size over time at different centers, blocked randomization could be used (26). To achieve an approximate balance between

the groups, regarding some important characteristics of the subjects, biased coin design, stratified randomization or minimization could be used (27–30).

Blinding

Clinical trials are often double-blind, meaning that neither the investigator nor the patient is aware whether the patient receives placebo or active treatment. The aim is to minimize subconscious effects, for instance optimistic judgements by the observers (observer bias). It is also well known that patients can experience benefits just by knowing or believing that they are on active treatment. This effect is also known as the placebo effect (31). One illustration is a study where the effect of ascorbic acid on the common cold was being investigated (32). The treatment showed a positive effect, but it was revealed that some of the subjects had opened the capsules and tasted the contents and thus became aware of receiving active drug or placebo. An analysis for taking into account the 'broken blindness' was performed and showed that there was no effect by this treatment.

Observational studies – case-control design

The different concepts described above are methodologically important especially for experimental studies. There are, however, a lot of situations where it is not possible to use an experimental design. To randomize groups into 'smoking' and 'not smoking' would not be possible for ethical reasons (nearly as impossible as randomizing gender). Thus to study whether variables like this are related to a specific health condition we must use an observational study. There are several possible designs (33). One of the most commonly used designs is the case-control study (34, 35). The weakness with such a design is that as many as 35 different possible sources of bias have been identified (36). A major difficulty is choosing an appropriate control group. The general principle is to choose patients as controls who might have been cases in the study (37). To follow this principle, it is common to use matched controls, by group or by individual. To use hospital controls, i.e. patients at the same hospital who are patients for other reasons, may be convenient but it may lead to an underestimated relationship between the possible risk factor and the health condition under study. Another alternative is to use a random sample of the general population as a control group. This is a good idea theoretically, but it may be difficult in practice, especially if we want a specific distribution regarding some variables, e.g. gender and age. In a case-control study it may be important to use blindness in the sense that the observer is blinded. If for instance a possible relationship between food intolerance and day/

night-work is investigated, it may be important that the diagnosis of food intolerance is made without knowing whether the subject is working in the day or at night. Blinding in this case is a protection against observer bias. The aim with a case-control study is often to study possible risk factors. A case-control study could be very valuable. However, due to the many possible biases, case-control studies should be interpreted with care (38). Furthermore, an association shown between a factor and a health condition in a case-control study is not in itself proof of causality. For claiming causality there are several other criteria to fulfil and it might be more useful to use the term 'risk indicator' instead of 'risk factor' as long as the causality is unproven (39–42).

In summary, an observational study, e.g. a case-control study, is in many situations the only possible design. It is important to evaluate possible biases and to conduct the study in a way that eliminates these biases as far as possible. This is an important and crucial aspect of the work for medical researcher and statistician jointly.

To choose a design

A controlled randomized double-blind study is often regarded as the gold standard and it is the most scientifically accepted design. It is clear that there is a certain hierarchy for different designs (43). Nevertheless, I believe that all designs could give valuable information if used in a careful manner. A recent study shows that the results could be comparable between different designs (44). To weigh up the general pros and cons for all possible designs is a good point to start at, giving consideration to statistical, medical, ethical, practical and financial aspects. Once the specific design has been chosen it is necessary to figure out how to minimize the influence of possible biases. Although this can be time-consuming it must be remembered that one single bias might be enough to invalidate the conclusion.

Statistical significance – rather an empty expression?

Assume that you challenge a colleague to 10 games of chess in a tournament. If you are equally good at chess we would expect the result to be 5–5, or close to it. Assume that the tournament ends up with a result of 3–7. There is no doubt that your colleague *was* better than you regarding this tournament – no inferential statistics is needed for that statement. But, the question is whether the result can be used to generalize. Could we make a switch in tense, from past tense to present tense? Your colleague *was* better, but *is* he/she better? To be able to answer that question we must analyse how likely it is to end up with a result which gives one of the players seven or more victories just by chance, i.e. given that the two players are equally good. As pointed out, we expect

the result to be 5–5 between two equally good players, but as a matter of fact there is a probability of 0.34 that one of the (equally good) players wins seven or more games just due to randomness. Thus, I would explain the result as a random outcome – certainly not proof of inferiority. However, if we change the scenario to a result of 1–9, which has a corresponding probability of 0.02, then it would be pitiable even for a bad loser (like myself) to explain that result as random or purely bad luck. In that case you would have to admit that your colleague *is* better.

In this example, the null-hypothesis is that the two players are equally good. For a specific result it is possible to calculate a *P*-value, which gives us the probability of achieving a result equally extreme or more extreme than our observed one, given that the null-hypothesis is true. For instance, if the result is 1–9 the *P*-value is 0.02, meaning that the probability of observing at least such a great difference (in any direction) between the two equally good players (null-hypothesis) is only 0.02. In medical research it is more or less standard to reject the null-hypothesis if the *P*-value is lower than 0.05. This means that the risk of wrongly rejecting the null-hypothesis equal to 5%, i.e. the level of significance is 5%.

Let us study another two examples. If the difference between two treatments is tested and the *P*-value is 0.02 it leads to the conclusion that the two treatments *differ* in effect. If a potential risk factor is studied in a case-control study and the *P*-value is below 0.05 we may conclude that there *is* a relationship (causality is not concluded). Observe that the conclusions are in present tense. The logic behind a hypothesis test may seem simple. However it is a concept which is subject to much misunderstanding and misinterpretation. To begin with the choice of 0.05 as a limit when it is 'beyond reasonable doubt' that the null-hypothesis is true is an arbitrary choice. If the *P*-value were 0.049, I would say that the proof value is not greater than if it were 0.051, but with 0.05 as a standard limit it is a difference between statistical significance or not. Actually it could be the difference between a 'significance star' or not in the article, or even worse, it could be the difference between published results or not. It is a fact that nonsignificant results are less likely to be published, i.e. publication bias (45, 46). Furthermore, a statistically significant effect of a treatment gives us no information about the magnitude of the effect and thus we cannot judge whether it is clinically relevant or not. Assume that diastolic blood pressure is being studied and two treatments are being compared. Imagine a study with a small estimated difference between the two treatments, e.g. 0.6 mmHg, and a narrow confidence interval with an absolute error of 0.2. This means that the true expected treatment difference is estimated to be between 0.4 and 0.8. Since zero is excluded the observed difference is statistically significant. But at the same time differences between 0.4 and 0.8 mmHg are not very clinically relevant, so the

two treatments could be considered as clinically equivalent in efficacy. If the result is not significant it is often misinterpreted as a proof of no clinical effect, but that is a wrong conclusion. Consider the chess tournament again: if the tournament ends with the result 3–7 it is not statistically significant. Thus we cannot conclude that your colleague *is* superior but we cannot conclude equality either. The only possible conclusion is that we do not have evidence enough for proofing a difference. In a survey of 71 published trials with no statistically significant results it was shown, by calculating confidence intervals instead of *P*-values, that nearly half of the trials showed a potential therapeutic improvement of 50% (47).

Thus a nonsignificant result does not exclude clinically relevant effects. This can be elegantly described by the phrase ‘Absence of evidence is not evidence of absence’ (48). Another misinterpretation about *P*-values is that *P*-values reflect the magnitude of the effect under study. It is a rather common belief that extremely low *P*-values, let us say lower than 0.001 implies a greater effect than in a studies with a *P*-value lower than 0.05. However, a low *P*-value can easily be achieved by using a large sample size. Extremely large studies can give small *P*-values even if the observed difference is small. Thus an extremely small *P*-value does not imply a great effect. The common misunderstandings presented above are the reasons for recommending confidence intervals instead of *P*-values in presentations (49). A confidence interval contains both information about the magnitude of the observed effect and information about whether the result is statistically significant. In summary, statistically significant results can be clinically relevant, or not clinically relevant, or even clinically equivalent (quite rare, but possible). Furthermore a nonsignificant result cannot exclude possible clinical efficacy. Thus, the expression ‘statistically significant’ is rather an empty expression. A confidence interval is the recommended presentation since it allows us to interpret statistical figures in a clinical perspective.

Sample size considerations

When testing hypotheses two things can go wrong. Firstly, we may reject the null-hypothesis even if it is true (a type I error). To use 0.05 as significance level means that the probability of making a type I error is 0.05. Secondly, we may accept the null-hypothesis even if it is false (a type II error). Instead of discussing the probability of making type II errors it is common to discuss the power of a test, which is the probability of correctly rejecting the null-hypothesis. In our tournament example the power is the probability to receive a significant result (concluding that one player is superior) given that one player actually is superior. If two treatments are to be compared, the power is the probability

of receiving significance if there is a true difference between the two treatments. Roughly, then, it is the chance of statistically proving a difference, if there actually is a difference.

It is obvious that the power is dependent of the magnitude of the difference. For instance, how likely is it that our chess tournament ends up with a significant result if your colleague is better? Well, it depends on the magnitude of the difference. If your colleague is among the top 10 players in the world, while you have only just learned the rules, the power is high even if only 10 games are played. But if your colleague is just slightly better than you, it is likely that this will be unproven (luckily for you) over only 10 games.

The power is also dependent on the level of significance, the variability of the variable studied, and the sample size. A large sample size could lead to a significant result even if the difference is small. But it is always relevant to raise this question if it is worth the effort to show a small difference. In the same way, it is not wise to use a sample size that is too small which could lead to a nonsignificant result even if the observed effect is of great clinical interest. As discussed earlier a survey of trials proved that nearly half of the nonsignificant trials showed a potential 50% therapeutic improvement. This implies that a lot of studies do not have large enough sample sizes. In another survey of trials it was concluded that most nonsignificant trials do not have samples of sufficient size for detecting relative differences as large as 25% or 50% between treatments (50).

Careful power considerations before the study starts are important. It is worth mentioning that power calculations (51, 52) in most cases rely on estimated difference and estimated variability, which implies that power and sample size calculations also are approximate. Thus, it may be a good idea to have a discussion regarding the observed result in relationship to the sample size used. For instance, the variability may be higher than expected and that could explain why the result is not statistically significant even if the estimated difference is clinically relevant and as large as expected.

Other dilemmas regarding statistical significance

If we study several different variables and perform a significance test on each of them, the risk of falsely rejecting at least one is fairly high. Assume that we perform a randomized study with two groups, and give both groups exactly the same treatment. If we analyse 14 independent variables, the probability of receiving at least one significance is more than 50%, even if we know that the treatments *de facto* are equal. Therefore, it is difficult to judge objectively whether a significant variable found by analysing a large number of variables is due to random effects or if it is due real effects. If a confirmatory trial includes several primary variables it is

possible to adjust for the multiple testing (53, 54). In an explanatory trial no hypothesis or primary variables need to be pre-defined. Such an approach can give plenty of valuable information. However, results from statistical significance analyses must be carefully interpreted. Due to the problem with multiple testing, a statistical significance may only be regarded as a generated hypothesis – a ‘flag’ telling us that there may be something interesting. Of course the proof value increases if the significance is medically motivated or if the same significance is found in other independent studies (confirmatory or explanatory).

Another more philosophical point is that the hypothesis ought to be stated in advance. Take the example of a small company of electricians where it is noticed that the proportion of boys among the employees’ children is extremely high (P -value 0.04). This could hardly be regarded as proof that electricians are more likely to have sons than daughters. If we consider all the companies in the world, there must be several with a high proportion of either boys or girls. In most situations we might expect a balance between boys and girls, and in such a company no one would think of testing whether there was a difference in proportion of boys and girls. Thus, to use a random finding to invent a hypothesis and to regard it as proven is a non-valid argument. But if the hypothesis (relationship between ‘electricity’ and gender) was generated in earlier trials and if this company was randomly chosen, then the significance would have a higher proof value. It is therefore important that the hypothesis and the corresponding variable are justified and stated in advance.

Concluding remarks

Firstly, I pointed out that it is important to have statistical input right from the planning phase of the study. Secondly, I indicated that there are a lot of

potential pitfalls at each step of the study. Thirdly, I explained that the different steps are dependent, that is the strength of the analysis depends on the design. These three arguments justify my conclusion that statistical issues are significantly important to medical research.

This discussion also justifies that statisticians should have an active part in the study team, not playing ‘second fiddle’ as an adviser on the odd occasion. Cooperation between clinicians and statisticians is essential to guarantee that the study is planned and performed in a way that allows meaningful analyses and valid conclusions. When a new study is scheduled I would like to recommend that the first meeting should not take place until a statistician is present; unfortunately there are too few statisticians available.

Nevertheless a lot of good information is available: there are excellent and comprehensive books written for readers who are nonstatisticians, for instance, the books written by Altman, Bland and Campbell et al. and Pocock are recommended (55–58). For statistical issues especially related to drug development, Senn has written an excellent book which also shows that statistics and humour are not mutually exclusive (59). There are also some guidelines available. With respect to the reporting of trials the CONSORT statement should be consulted (60). A detailed guideline which also contains checklists used by several medical journals is given in a book edited by Gardner and Altman (61). In order to perform meta-analyses (62) it must be possible to judge the details of the study design; for this reason reporting a trial according to guidelines is important.

Even if some answers are given in the literature, I hope that the future will bring increased cooperation between statisticians and medical researchers. It would benefit the development of both medical research and statistical research, it would benefit the quality of science, and it would benefit the patients.

References

1. ALTMAN DG. Statistics in medical journals. *Statistics Med* 1982;1:59–71.
2. ALTMAN DG. Statistics in medical journals. Development in the 1980s. *Statistics Med* 1991;10:1897–1913.
3. Editorial: The scandal of poor medical research. *Br Med J* 1994;308:283–284.
4. EMERSON JD, COLDITZ GA. Use of statistical analysis in the New England Journal of Medicine. *N Engl J Med* 1983;309:709–713.
5. SCHOOLMAN HM, BECKTEL JM, BEST WR, JOHNSON AF. Statistics in medical research: principals versus practice. *J Laboratory Clin Med* 1968;71:357–367.
6. NOLLER KL, MELTON LJ. Study design in perinatal medicine. *Am J Perinatol* 1985;2:250–255.
7. SACKET DL. Rational therapy in the neurosciences: the role of the randomized trial. *Stroke* 1986;17:1323–1329.
8. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use., Guidelines: Statistical Principles for Clinical Trials (ICH E-9). Available from URL: <http://www.ifpma.org/ich5e.html#Design>.
9. CAMPBELL MJ, MACHIN D. Medical Statistics. A commonsense approach, 3rd edn. Chichester: Wiley, 1999, 28–31.
10. FLEMING T. Surrogate endpoints in clinical trials. *Drug Information J* 1996;30:545–551.
11. PRENTICE RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics Med* 1989;8:431–440.
12. McDOWELL I, NEWELL C. Measuring Health. A Guide to Rating Scales and Questionnaires. Oxford: Oxford University Press, 1987.
13. STREINER DL, NORMAN GR. Health Measurement Scales. A Practical Guide to Their Development and Use. Oxford: Oxford University Press, 1989.
14. SENN S. Statistical Issues in Drug Development. Chichester: Wiley, 1997, 28–29.
15. BLAND M, ALTMAN DG. Regression towards the mean. *Br Med J* 1994;308:1499.

16. BLAND M, ALTMAN DG. Some examples of regression towards the mean. *Br Med J* 1994;**309**:780.
17. CHUANG-STEIN C. The regression fallacy. *Drug Information J* 1993;**27**:1213–1220.
18. ARMITAGE P, FOX W, ROSE GA, TINKER CM. The variability of measurements of casual blood pressure II. Survey experience. *Clin Sci* 1966;**30**:337–344.
19. YUDKIN P, STRATTON IM. How to deal with regression to the mean in intervention studies. *Lancet* 1996;**347**:241–243.
20. HAYES RJ. Methods for assessing whether change depends on initial value. *Statistics Med* 1988;**7**:915–927.
21. POCKOCK SJ. Randomised clinical trials. *Br Med J* 1977;**i**:1661.
22. SACKS HS, CHALMERS TC, SMITH H. Sensitivity and specificity of clinical trials: randomized v historical controls. *Arch Intern Med* 1983;**143**:753–755.
23. CHALMERS TC, CELANO P, SACKS HS, SMITH H. Bias in treatment assignment in controlled clinical trials. *New Eng J Med* 1983;**309**:1358–1361.
24. SCHULZ KF, CHALMERS I, HAYES RJ, ALTMAN DG. Bias due to non-concealment of randomization and non-double-blinding. *J Am Med Association* 1995;**273**:408–412.
25. ROTHMAN KJ. Statistics in nonrandomized studies. *Epidemiology* 1990;**1**:417–418.
26. POCKOCK SJ. *Clinical Trials*. Chichester: Wiley 1983, 73–89.
27. EFRON B. Forcing a sequential experiment to be balanced. *Biometrika* 1971;**58**:403–417.
28. ATKINSON A. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* 1982;**69**:61–67.
29. COCHRAN WG. *Sampling Techniques*. New York: Wiley 1977.
30. TAVES DR. Minimization: a new method of assigning patients to treatment and control groups. *Clin Pharmacol Therapeutics* 1974;**15**:443–453.
31. BYERLY H. Explaining and exploiting placebo effects. *Pers Biol Med* 1976;**19**:423–435.
32. KARLOWSKI TR, CHALMERS TC, FRENKEL LD et al. Ascorbic acid for the common cold. A prophylactic and therapeutic trial. *JAMA* 1975;**231**:1038–1042.
33. BAILAR JC, LOUIS TA, LAVORI PW, POLANSKY M. A classification for biomedical research reports. *N Engl J Med* 1984;**311**:1482–1487.
34. BRESLOW NE, DAY NE. *Statistical Methods in Cancer Research, Vol. 1. –the Analysis of Case-Control Studies*. Oxford University Press, 1993.
35. SCHLESSELMAN JJ. *Case-control studies. Design Conduct, Analysis*. Oxford University Press 1982.
36. SACKET DL. Bias in analytical research. *J Chron Dis* 1979;**32**:51–63.
37. ROTHMAN KJ, GREENLAND S. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven 1998.
38. MAYES LC, HORWITZ RI, FEINSTEIN AR. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol* 1988;**17**:680–685.
39. BECK JD. Risk revisited. *Community Dent Oral Epidemiol* 1998;**26**:220–225.
40. KLEINBAUM D, KUPPER L, MORGENTERN H. *Epidemiologic research: principals and quantitative methods*. Belmont, CA: Lifetime Learning Publications 1982.
41. HILL B. *Principals of medical statistics*, 9th edn. New York: Oxford University Press 1971, 309–323.
42. KOCH GG, BECK JD. Statistical methodologies useful for the analysis of data from risk-assessment studies. *J Public Health Dent* 1992;**52**:146–167.
43. CONCATO J, SHAH N, HORWITZ RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–1892.
44. BENSON K, HARTZ AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–1886.
45. EASTERBROOK PJ, BERLIN JA. Publication bias in clinical research. *Lancet* 1991;**337**:867–872.
46. BERLIN JA, COLIN BB, LOUIS TA. An assessment of publication bias using a sample of published clinical trials. *J Am Stat Association* 1989;**84**:381–392.
47. FREIMAN JA, CHALMERS TC, SMITH H, KUEBLER RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of 71 'negative' trials. *N Engl J Med* 1978;**299**:690–694.
48. ALTMAN DG, BLAND M. Absence of evidence is not evidence of absence. *Br Med J* 1995;**311**:485.
49. GARDNER MJ, ALTMAN DG. Confidence intervals rather than P-values: estimation rather than hypothesis testing. *Br Med J* 1986;**292**:746–750.
50. MOHER D, DULBERG CS, WELLS GA. statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;**272**:122–124.
51. MACHIN D, CAMPBELL MJ, FAYERS PM, PINOL APY. *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Scientific 1997.
52. DESU MM, RHAGHAVARAO D. *Sample Size Methodology*. Boston: Academic Press 1990.
53. BAUER P. Multiple testing in clinical trials. *Stat Med* 1991;**10**:871–890.
54. POCKOCK SJ, GELLER N, TSIATIS A. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;**43**:487–498.
55. ALTMAN DG. *Practical Statistics for Medical Research*. London: Chapman & Hall 1991.
56. BLAND M. *An introduction to medical statistics*, 3rd edn. Oxford: Oxford University Press 2000.
57. CAMPBELL MJ, MACHIN D. *Medical Statistics: A commonsense approach*, 3rd edn. Chichester: Wiley 1999.
58. POCKOCK SJ. *Clinical Trials*. Chichester: Wiley 1983.
59. SENN S. *Statistical Issues in Drug Development*. Chichester: Wiley 1997.
60. BEGG C, CHO M, EASTWOOD S et al. improving the quality of reporting of randomized controlled trials. CONSORT Statement. *JAMA* 1996;**276**:637–639. (Also available at url: <http://www.consort-statement.org>.)
61. GARDNER MJ, ALTMAN DG, , eds. *Statistics, with confidence. Confidence intervals and statistical guidelines*. London: British Medical Journal 1989.
62. SACKS HS, BERRIER J, REITMAN D, ANCONA-BERK VA, CHALMERS TC. Meta-analyses of randomized controlled trials. *New Eng J Med* 1987;**316**:450–455.