Statistics and medical research

Basic principles review

Jonas Ranstam

Department of Community Medicine, Lund University, and the NEPI Foundation, Malmö General Hospital, S-214 01 Malmö, Sweden. Tel +46 40-332664. Fax -336215

Modern medical research often includes too many statistics which are misused or misunderstood. This problem has been identified and discussed by several authors, see, e.g., Altman et al. 1983, Editorial 1988, 1992, Morris 1988. In this article I will explain some general reasons for applying statistical principles and give a few hints about how to do this. The interested reader can find detailed information on the interpretation of specific statistical methods in modern textbooks, e.g., Machin and Campbell (1993), Pagano and Gavreau (1993).

The need to employ statistical principles in a project depends on the type of study that is to be undertaken. A description of a clinically interesting case (a case study) may be presented with few, if any, statistical aspects in mind, whereas a project comparing cases with a disease and healthy controls (a case-control study) to assess specific risk factors (common to all of us), or outcome of different treatments, may need the full collaboration of a professional statistician. Most investigators, however, need some insight into statistical principles, which are an integrated part of modern medical research methodology. This is important for the communication of quantitative results since the risk of confusion will be reduced if the principles are applied. But the aim of research methodology is much more than that, i.e., valid and reliable conclusions.

Design

At its best, a scientific paper presents a clear and well structured story. A common feature of clinical research is often a specific agent to which subjects have been exposed; a risk or a prognostic factor, a surgical procedure or instrument, a prosthesis, or a chemotherapy. This exposure is measured in terms of effects on wellbeing, stage of disease, longevity, etc. For these effects to be measured empirically, a study design must be created. This creation links the conceptual aim of the study with a specific operational hypothesis that either can be tested empirically in a prospective randomized hypothesis testing study or, if the study population was defined for some other purpose, can be used as the framework for description of empirical data in a hypothesis-generating study. In the latter case, the new hypotheses should be validated in a new prospective study.

For instance, to investigate if calcium supplementation can reduce the risk of a Colles' fracture, a study population or sample has to be defined and observed during a specified time period. A standardized scheme of calcium supplementation with respect to age at start, dosage, duration of treatment, etc., should be established and a control group without calcium supplementation should be selected. A method for detecting the fractures in the studied population must be devised, the information must be recorded and computerized, etc.

A mistake when designing a study may be detrimental to the conclusions. If persons suffering from osteoporosis, in the studied population, have been prescribed calcium supplementation as a preventive measure and this is not taken into account (when the preventive effect of calcium is studied), the calcium exposed subjects will appear to constitute a high-risk group for fractures. An impeccable design, however, is not enough for drawing correct conclusions from a study.

Validity and precision

When considering whether information obtained from a restricted sample of patients in a clinical study applies to all individuals in the general population, the cardinal issues are validity and precision. Many investigators seem to rely only on the statistical significance when trying to assess this question. The

broader aim of verifying that the information is valid as well as not due to random variation is much more important than a mere search for statistical significance.

Validity concerns systematic measurement errors. Bias is a synonym for such errors.

Validity

Bias can emerge in one or more of 3 ways:

1) Selection problems exist if the studied patients are not representative of the general population. When studying a series of consecutive patients some may decline participation in the study. Participation rate per se is, however, not as important as the representativity. As long as the patients who decline participation do not differ systematically from others it is of little consequence.

Selection bias will occur if they do differ systematically. If, in the example of Colles' fractures and calcium supplementation, all persons with previous fragility fractures decline to participate in the study, the estimate of the effect from calcium supplementation may not be valid, i.e., the effects would not be the same if calcium supplementation had been carried out in the general population. In most cases, selection bias cannot be adjusted for in the statistical analysis.

2) Information from study subjects may be wrongly classified. The reliability, or repeatability, of an instrument for collecting information is often assessed using repeated measurements in a subsample of the studied population and in terms of a reliability coefficient, e.g., kappa. The overall quality of the instrument is generally measured in terms of sensitivity and specificity.

In some situations, misclassification leads to an attenuation of a studied relationship. This could be the case, in our example, if some of the control subjects used calcium in spite of being controls. Artifactual relationships can, of course, be generated by the same mechanisms. Information problems should be anticipated and prevented in the design of the study by considering alternative or complementary instruments for the collection of information or by adjusting the sample size.

3) Confounding is a problem that sometimes can be solved in the statistical analysis. This type of bias can be thought of as a distortion in the estimate of the exposure effect that arises because cases and controls are not comparable with respect to some important background factor, e.g., age, gender or smoking habits.

Suppose, in the example, that the calcium users who participated in the study were younger than their controls, and that the age difference was not considered in the analysis. As age is a risk factor, calcium use will, other things being equal, appear more related to a reduction in fracture-risk than it actually is.

Potential confounding factors can be incorporated in stratified or multivariate analyses and be evaluated simultaneously with the investigated exposure effect. The confounding bias may then be dealt with. It is, however, a practical problem that potential confounding factors should be considered already in the design stage so that data on them are available in the analysis stage.

Precision

Random variation, unlike bias, can be evaluated using statistical tests. A statistical test results in a p-value with information on the risk of drawing a false-positive conclusion. It is a generally accepted principle that this risk (the significance level) should not be greater than 5 percent. This introduces an artificial dichotomy into statistically significant and not significant that often is too simplistic.

Suppose that the fracture rate for calcium users turns out to be 23 percent lower than the rate among the controls. Is the difference in rates large enough not to be explained simply by random variation? Using a statistical test, e.g., a Chi-square test or a Fischer's exact test, a p-value can be calculated. If the p-value is 12 percent, then the risk of drawing a false-positive conclusion, i.e., that the exposure to calcium supplementation reduces the fracture rate, is greater than the allowable 5 percent. Hence, we refrain from concluding that we have observed such an effect.

A test that is not statistically significant, however, does not indicate that the exposure has no effect. In the example, we cannot conclude that calcium supplementation is worthless. Our conclusion is simply that we cannot detect any effect. The p-value should not be confused with the ability to avoid a false-negative conclusion which is measured by the statistical power, not by the p-value. Non-significant test results thus convey adequate information only if the statistical power is known.

A study launched to give a scientific answer to a specific question should, in general, not be designed with a statistical power lower than 80 percent. That aim is achieved by including a sufficient number of subjects in the study. The exact number of subjects needed to answer the question can easily be found in a sample-size table, e.g., Machin and Campbell (1987), Lwanga and Lemeshow (1991), or by consulting a statistician.

Since the p-value contains no information on the statistical power, but often conveys a false impression of total statistical precision, it is sometimes regarded

as inferior to the 95 percent confidence interval. The width of this interval, in some sense, reveals statistical power, or the margin of error, at the same time as its limits can indicate statistical significance, i.e., if a confidence interval for the ratio of fracture rates, between subjects given calcium supplements and their controls, is (95% CI: 0.8–1.1) and thus includes the value 1.0, then an effect of calcium supplementation would not be statistically significant. The value 1.0 is, of course, equivalent to identical fracture rates in the two groups. Had the confidence interval instead excluded the possibility of identical rates, e.g., been (95% CI: 0.6–0.9) then the effect of calcium supplementation would have been significant and the corresponding p-value lower than 5 percent.

How statistical principles should be used in a manuscript

The *introduction* section should clearly present the conceptual aims of the study.

The *material* (patients, animals) *and methods* section is the place for operational definitions. The study design and the studied population(s) should be presented here. It should include statements on why the used sample size was chosen. It should clearly describe how the subjects were selected and what potential selection mechanisms could have affected the validity of the data. The treatment, surgical procedure or epidemiological exposure under study should be adequately defined in technical terms. The means and methods of collecting information should be presented and the tests used in the statistical analysis should be clearly defined and referenced if unusual.

The *results* section should present all relevant data as clearly as possible and whenever a statistical test has been used the reader should be able to identify what specific test it was. If possible the reader should be allowed to judge for himself if the assumptions for the test were fulfilled, i.e., by presenting auxiliary information, if necessary.

The precision (statistical variability) of the data should accompany the results. This can be described using confidence intervals or standard errors, but sometimes a p-value may be appropriate. Never present a p-value without describing the quantity tested. Avoid using dichotomies (* and NS) instead of p-values.

The *discussion* section is the proper place for reasoning about validity. Could a selection or misclassification have biased the results? If a certain bias does exist, could it be quantified? If not, would it be possible to assess whether or not it leads to an overestimation of the true effect? Could confounding bias be a real problem? Have alternative explanations for the findings been investigated?

The *conclusions* should be consistent with both the results section and the discussion section.

Since medical papers usually are not written for statisticians, the language in the manuscript should, apart from the material and methods sections, be as precise and easy to understand as possible. A manuscript does not benefit from unnecessary logical complexity or presentation of intermediate statistical results. Avoid using statistical terms when they are not needed, e.g., do not use the words correlation instead of relation, covariate of age instead of age, odds ratio instead of relative risk, etc. Always assess the clinical significance of statistically significant differences. A difference between 10% and 15% patients cured may, for instance, be statistically significant but clinically irrelevant.

Lastly, remember that statistics is more a scientific method for drawing conclusions from incomplete information than a tool for proving the correctness of a finding.

References

Altman D G, Gore S M, Gardner M J, Pocock S J. Statistical guidelines for contributors to medical journals. Br Med J 1983; 286: 1489–93.

Editorial. Art—or science? J Bone Joint Surg (Br) 1988; 70: 173.

Editorial. Statistics in The Journal of Bone and Joint Surgery: Suggestions for authors. J Bone Joint Surg (Am) 1992; 74: 319–20.

Lwanga, S K, Lemeshow S. Sample size determination in health studies. A practical manual. WHO, Geneva 1991.

Machin D, Campbell M. Statistical tables for the design of clinical trials. Blackwell, Oxford 1987.

Machin D, Campbell M. Medical statistics. Wiley, New York 1993.

Morris R. A statistical study of papers in the Journal of Bone and Joint Surgery (Br) 1984. J Bone Joint Surg (Br) 1988; 70: 242–6.

Pagano M, Gavreau K. Principles of Biostatistics. Duxbury press, Belmont, California 1993.