ENDGAMES

STATISTICAL QUESTION

Multiple hypothesis testing and Bonferroni's correction

Philip Sedgwick *reader in medical statistics and medical education*

Institute for Medical and Biomedical Education, St George's, University of London, London, UK

Researchers assessed the effectiveness of a multifaceted intervention directed at general practitioners on six year mortality and morbidity in patients with newly diagnosed type 2 diabetes. A cluster randomised controlled study design was used. Clustering was at the GP level. The intervention consisted of regular follow-up and individualised goal setting for patients, which was supported by prompting of doctors, clinical guidelines, feedback, and continuing medical education. The control treatment consisted of routine care, and doctors were free to choose any treatment and change it over time. Participants were aged 40 years or more, had been diagnosed as having diabetes during 1989 to 1991, and had survived until six year follow-up. In total, 874 patients were recruited, with 459 allocated to the intervention and 415 allocated to the control.[1]

The primary outcomes were overall mortality, incidence of diabetic retinopathy, urinary albumin concentration ≥15 mg/L, myocardial infarction, and stroke in patients without these outcomes at baseline. The critical level of significance was 0.05 (5%). At the end of follow-up, the treatment groups differed significantly only in one of the five primary outcomes. A lower proportion of the intervention group had a urinary albumin concentration ≥15 mg/L (22.5% *v* 30.8%; P=0.04). When multiple testing was taken into account using Bonferroni's adjustment, no significant differences in the primary outcomes were observed. It was concluded that, in primary care, individualised goals with educational and surveillance support did not affect six year mortality and morbidity in patients with newly diagnosed type 2 diabetes.

Which of the following statements, if any, are true?

a) The maximum probability of a type I error occurring for a single hypothesis test was 0.05 (5%)

b) The overall type I error rate for the five hypothesis tests of the primary outcomes was no greater than 0.05 (5%)

c) Bonferroni's correction was used to reduce the probability of a type I error occurring when multiple testing

d) The adjusted critical level of significance using Bonferroni's correction for each primary outcome was 0.01 (1%)

Answers

Statements *a*, *c*, and *d* are true, whereas *b* is false.

The aim of the trial was to investigate the effectiveness of a multifaceted intervention directed at GPs on six year mortality and morbidity in patients with newly diagnosed type 2 diabetes. A randomised controlled trial study was performed. The treatment groups were compared in five primary outcomes that measured mortality and morbidity. Traditional statistical hypothesis testing, described in a previous question,[2] was used to compare treatment groups in each outcome. For each outcome the test of hypotheses started at the position of equipoise. The null hypothesis stated that, in the population of newly diagnosed patients with type 2 diabetes from which the sample was obtained, no difference existed between treatment groups at six year follow-up.

For each statistical test, it was possible that the comparison of treatment groups resulted in a type I error. A type I error, described in a previous question,[3] would have occurred if the null hypothesis was rejected in favour of the alternative—that is, a significant result was seen even though no difference existed between treatment groups in the population. The critical level of significance for a hypothesis test was 0.05 (5%). Therefore, the probability of a significant difference occurring for a hypothesis test was 0.05 (5%). Hence the maximum probability of a type I error occurring for a single test was 0.05 (5%) (*a* is true), termed the type I error rate. However, five hypothesis tests were performed in the above study. As described in a previous question,[4] because multiple testing was performed the probability of any one of these tests resulting in a significant difference was greater than 0.05 (5%). Therefore, the overall type I error rate was greater than 0.05 (5%) (*b* is false).

The probability that at least one of the five hypothesis tests of the primary outcomes in the above study would result in a significant difference is relatively straightforward to obtain. The probability that one of the tests will not be significant was 0.95. If it is assumed that the five hypothesis tests were independent of each other (outcomes being tested were not correlated with each other), using the multiplicative rule for the occurrence of independent events, the probability that all five will not be

p.sedgwick@sgul.ac.uk

significant is 0.95^5. The probability that at least one of the hypothesis tests will be significant is therefore $(1-0.95^5)=0.2262$. Thus, the type I error rate after multiple testing of the primary outcomes for the above study was 0.2262 (*b* is false). More generally, the probability that at least one hypothesis test will be significant when multiple testing is $(1-0.95^x)$, where x is the number of statistical tests performed. Therefore, as the number of statistical tests (x) increases, the value of 0.95^x decreases, and the probability of a significant difference and a type I error also increases.

When undertaking multiple hypothesis testing of the primary outcomes in the study above, it was essential that the type I error rate was reduced and kept at approximately the original critical level of significance—that is, 0.05 (5%). To achieve this, the critical level of significance for each test would need to be adjusted and be smaller than 0.05 (5%). It is relatively straightforward to obtain the new critical level of significance for each hypothesis test needed to control the overall type I error rate. The new critical level of significance (denoted by α) would be equal for each of the five tests. The critical level of significance is the probability that a single test will be significant. Therefore, the probability that a single test will not be significant is $(1-\alpha)$. If it is assumed that the tests were independent of each other, the probability that all five hypothesis tests will not be significant is $(1-\alpha)^5$. To maintain a type I error rate of 5% when undertaking the five hypothesis tests, $(1-\alpha)^5$ must equal 0.95. Because α is small, it can be shown that $(1-\alpha)^5$ is roughly equal to $(1-5\alpha)$. The details of this approximation are beyond the scope of this article. For $(1-5\alpha)$ to equal 0.95, then 5α must equal 0.05—that is, $\alpha=0.05\div5=0.01$. Therefore, the adjusted critical level of significance needed to maintain an overall type I error rate of 5% for the testing of the five primary outcomes in the above study was 0.01 (1%). This is the basis for Bonferroni's correction. More generally, Bonferroni's correction adjusts the critical level of significance for each test by dividing the critical level of significance, typically 0.05 (5%), by the number of significance tests performed. The aim of the correction is to maintain the type I error rate at about 5% and thereby reduce the probability of a type I error occurring when multiple testing (*c* is true).

For the above study, the researchers undertook five significance tests for the primary outcomes and, as described, the new critical level of significance after Bonferroni's correction was $0.05\div5=0.01$ (1%) (*d* is true). Before the adjustment for multiple testing, the only comparison of the treatment groups in the primary outcomes that was significant was that for urinary albumin concentration ≥15 mg/L (P=0.04). When multiple testing was taken into account using Bonferroni's correction, no significant differences were seen. Therefore, on the basis of Bonferroni's correction, it would seem reasonable to infer, although it cannot be confirmed, that the significant result originally seen between treatment groups in urinary albumin concentration ≥15 mg/L was a type I error.

Bonferroni's correction provides a straightforward approach to controlling the type I error rate when multiple testing is performed. It is appropriate when the number of tests is small. However, the correction tends to be conservative if a large number of tests are performed. In particular, the adjustment provides a new critical level of significance for each test that is smaller than needed to maintain the type I error rate at 5% overall when multiple testing. The consequence of the correction being conservative is that for any hypothesis test the null hypothesis may not be rejected in favour of the alternative despite there being a difference in the population. Therefore, Bonferroni's correction errs on the side of non-significance and its use will inflate the type II error rate. A type II error would occur if the null hypothesis was not rejected in favour of the alternative—that is, a non-significant result is seen even though a difference exists in the population. Furthermore, Bonferroni's correction does not adjust the extent to which outcomes may be correlated, which, if substantial, leads to conservative corrections.

Bonferroni's correction is not without its opponents, who claim that it is contrary to sensible scientific inference. If comparison between groups is based on the P value alone, then the correction will result in any given comparison being interpreted differently depending on the number of statistical tests performed. Researchers vary in the number of comparisons that are included when using Bonferroni's correction. Presumably it would be sensible to count all the statistical tests presented in a publication. However, it might be more appropriate to count all the tests that were performed, including those not published. For the example above, the researchers presented more than 100 statistical tests in the original publication. However, Bonferroni's correction was applied only to the tests for the five primary outcomes. It is not uncommon for researchers to restrict the number of statistical tests that Bonferroni's correction is applied to in a publication. This may be in an attempt to avoid implementing too conservative a correction. Nonetheless, it has been suggested that the number of statistical tests undertaken is irrelevant and that it may be unnecessary to adjust for multiple testing. Presumably, scientific inference should be based on what evidence the data provide, not statistical significance and how many tests were performed. In that respect, it might be that clinical significance should play a greater role in the inference of treatment effectiveness.[5] Finally, Bonferroni's correction is conservative and inflates the type II error rate. Presumably the consequences of a type II error are no less serious than those of a type I error.

Care must be taken when interpreting the results of studies that incorporate a large number of statistical tests to compare treatment groups. Ultimately, multiple testing will result in type I errors. However, it is not possible to ascertain which significant findings are a type I error. Bonferroni's correction is the most widely used approach for controlling the type I error rate. Although Bonferroni's correction is conservative, it does avoid spurious significant results. Other approaches have been developed, including the Holm method, which will be described in a future question. Although it is argued that Holm's method is superior to Bonferroni's correction because it is less conservative, there is still debate as to whether any adjustment should be made for multiple testing.

Problems with multiple testing also occur in types of analysis other than the comparison of treatment groups in multiple outcomes. Sometimes patients have several measurements of an outcome over time—for example, weight during pregnancy. It would be inappropriate to compare groups of patients at each time point with the application of Bonferroni's correction because the multiple observations for each patient would be correlated and Bonferroni's correction subsequently highly conservative. Approaches to analysing such data have been described in a previous question.[6]

Competing interests: None declared.

1 De Fine Olivarius N, Beck-Nielsen H, Andreasen AH, Hørder M, Pedersen PA. Randomised controlled trial of structured personal care of type 2 diabetes mellitus. *BMJ* 2001;323:970.
2 Sedgwick P. Understanding statistical hypothesis testing. *BMJ* 2014;348:g3557.
3 Sedgwick P. Pitfalls of statistical hypothesis testing: type I and type II errors. *BMJ* 2014;349:g4287.
4 Sedgwick P. Pitfalls of statistical hypothesis testing: multiple testing. *BMJ* 2014;349:g5310.
5 Sedgwick P. Clinical significance versus statistical significance. *BMJ* 2014;348:g2130.

6 Sedgwick P, Marston L. Analysis of longitudinal studies. *BMJ* 2013;346:f363.

Cite this as: *BMJ* 2014;349:g6284