

# What's wrong with Bonferroni adjustments

Thomas V Perneger

Institute of Social  
and Preventive  
Medicine,  
University of  
Geneva, CH-1211  
Geneva 4,  
Switzerland  
Thomas V  
Perneger,  
*medical epidemiologist*

Correspondence to:  
Dr Perneger  
pernegert@cmu.  
unige.ch

*BMJ* 1998;316:1236-8

When more than one statistical test is performed in analysing the data from a clinical study, some statisticians and journal editors demand that a more stringent criterion be used for "statistical significance" than the conventional  $P < 0.05$ .<sup>1</sup> Many well meaning researchers, eager for methodological rigour, comply without fully grasping what is at stake. Recently, adjustments for multiple tests (or Bonferroni adjustments) have found their way into introductory texts on medical statistics, which has increased their apparent legitimacy.<sup>2-3</sup> This paper advances the view, widely held by epidemiologists, that Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference.<sup>4-5</sup>

## Adjustment for multiple tests

Bonferroni adjustments are based on the following reasoning.<sup>1-3</sup> If a null hypothesis is true (for instance, two treatment groups in a randomised trial do not differ in terms of cure rates), a significant difference ( $P < 0.05$ ) will be observed by chance once in 20 trials. This is the type I error, or  $\alpha$ . When 20 independent tests are performed (for example, study groups are compared with regard to 20 unrelated variables) and the null hypothesis holds for all 20 comparisons, the chance of at least one test being significant is no longer 0.05, but 0.64. The formula for the error rate across the study is  $1 - (1 - \alpha)^n$ , where  $n$  is the number of tests performed. However, the Bonferroni adjustment deflates the  $\alpha$  applied to each, so the study-wide error rate remains at 0.05. The adjusted significance level is  $1 - (1 - \alpha)^{1/n}$  (in this case 0.00256), often approximated by  $\alpha/n$  (here 0.0025). What is wrong with this statistical approach?

## Problems

### Irrelevant null hypothesis

The first problem is that Bonferroni adjustments are concerned with the wrong hypothesis.<sup>4-6</sup> The study-wide error rate applies only to the hypothesis that the two groups are identical on all 20 variables (the universal null hypothesis). If one or more of the 20  $P$  values is less than 0.00256, the universal null hypothesis is rejected. We can say that the two groups are not equal for all 20 variables, but we cannot say which, or even how many, variables differ. Such information is usually of no interest to the researcher, who wants to assess each variable in its own right. A clinical equivalent would be the case of a doctor who orders 20 different laboratory tests for a patient, only to be told that some are abnormal, without further detail. Thus, Bonferroni adjustments provide a correct answer to a largely irrelevant question.

### Inference defies common sense

Bonferroni adjustments imply that a given comparison will be interpreted differently according to how many other tests were performed. For example, the

## Summary points

Adjusting statistical significance for the number of tests that have been performed on study data—the Bonferroni method—creates more problems than it solves

The Bonferroni method is concerned with the general null hypothesis (that all null hypotheses are true simultaneously), which is rarely of interest or use to researchers

The main weakness is that the interpretation of a finding depends on the number of other tests performed

The likelihood of type II errors is also increased, so that truly important differences are deemed non-significant

Simply describing what tests of significance have been performed, and why, is generally the best way of dealing with multiple comparisons

difference in remission rates between two chemotherapeutic treatments could be interpreted as statistically significant or not depending on whether or not survival rates, quality of life scores, and complication rates were also tested. In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be deemed normal. Surely this is absurd, at least within the current scientific paradigm. Evidence in data is what the data say—other considerations, such as how many other tests were performed, are irrelevant.

### Increase in type II errors

Type I errors cannot decrease (the whole point of Bonferroni adjustments) without inflating type II errors (the probability of accepting the null hypothesis when the alternative is true).<sup>4</sup> And type II errors are no less false than type I errors. In clinical practice, if a high concentration of creatine kinase were considered compatible with "no myocardial infarction" by virtue of a Bonferroni adjustment, the patient would be denied appropriate care. In research, an effective treatment may be deemed no better than placebo. Thus, contrary to what some researchers believe, Bonferroni adjustments do not guarantee a "prudent" interpretation of results.

### What tests should be included?

Most proponents of the Bonferroni method would count at least all the statistical tests in a given report as a basis for adjusting  $P$  values. But how about tests that were performed, but not published, or tests published in other papers based on the same study? If several

papers are planned, should future ones be accounted for in the first publication? Should we worry about error rates related to an investigator—taking the number of tests he or she has done in their lifetime into consideration<sup>6</sup>—or error rates related to journals? Should confidence intervals, which are not statistical tests, but are often interpreted as such (the confidence interval includes 0, hence the groups do not differ) be counted? No statistical theory provides answers for these practical issues.

### A futuristic scenario

What would happen to biomedical research if Bonferroni adjustments became routine? Cynical researchers would slice their results like salami, publishing one P value at a time to escape the wrath of the statistical reviewer. Idealists would conduct studies to examine only one association at a time—wasting time, energy, and public money. Meta-analysts would go out of business, since a pooled analysis would invalidate retrospectively all original findings by adding more tests to be adjusted for. Journals would have to create a new section entitled “P value updates,” in which P values of previously published papers would be corrected for newly published tests based on the same study. And so on ....

### Back to the Neyman-Pearson theory

These objections seem so compelling that the reader may wonder why adjustments for multiple tests were developed at all. The answer is that such adjustments are correct in the original framework of statistical test theory, proposed by Neyman and Pearson in the 1920s.<sup>7</sup> This theory was intended to aid decisions in repetitive situations. Imagine that your factory produces light bulbs in lots of 1000, and that testing each bulb before shipment would be impractical. You can decide to test only a sample in each lot, and to reject

(literally) any lots in which more than a predefined number ( $x$ ) of bulbs in the sample are defective. Of course, your decision might be wrong for any particular lot, but the Neyman-Pearson theory provides a decision rule (the number  $x$ ), so that over many trials your error rates (type I and type II) will be minimised. Now, if for some reason you took 20 samples out of a given lot instead of one, and decided that you would reject the lot if the number of defective bulbs exceeded  $x$  in only one sample, you would be much too likely to reject a good lot in error, and a Bonferroni adjustment would restore the original optimal error rates.

The catch is that Neyman and Pearson developed their statistical tests to aid decision making, not to assess evidence in data. The latter practice may be objected to for several reasons (this topic would deserve a discussion of its own), and alternative approaches to statistical inference, such as estimation procedures, use of likelihood ratios, and Bayesian methods, have been proposed.<sup>8–11</sup> Bonferroni adjustments follow the original logic of statistical tests as supports of repeated decisions, but they are of little help in determining what the data say in one particular study.

### Should Bonferroni adjustments ever be used?

Statistical adjustment for multiple tests make sense in a few situations. Firstly, the universal null hypothesis is occasionally of interest. For instance, to verify that a disease is not associated with an HLA phenotype, we may compare available HLA antigens (perhaps 40) in a group of cases and controls. If no association existed, at least one test would be significant with a probability of 0.87, and Bonferroni adjustments would protect against making excessive claims. A clinical equivalent is the case of a healthy person undergoing several laboratory tests as part of a general health check. Secondly, adjustments are appropriate when the same test is repeated in many subsamples, such as when stratified analyses (by age group, sex, income status, etc) are conducted without an a priori hypothesis that the primary association should differ between these subgroups. Note that this is the scenario, reminiscent of repeated sampling of the same lot, that Tukey and Bland and Altman use in their justifications of multiple test adjustments.<sup>1–3</sup> Sequential testing of trial results also falls in this category. A final situation in which Bonferroni adjustments may be acceptable is when searching for significant associations without pre-established hypotheses.

### The best approach

However, even in these situations, simply describing what was done and why, and discussing the possible interpretations of each result, should enable the reader to reach a reasonable conclusion without the help of Bonferroni adjustments.<sup>5–12</sup> There is an important difference between what the data say and what the researcher (or the reader) believes to be true.<sup>8</sup> The latter depends not only on the data at hand but also on considerations such as whether a finding is biologically plausible or whether the significant test was a serendipitous finding in a fishing expedition. The



DAVID HITCH

integration of prior beliefs with evidence is best achieved by Bayesian methods, not by Bonferroni adjustments. In summary, Bonferroni adjustments have, at best, limited applications in biomedical research, and should not be used when assessing evidence about specific hypotheses.

I thank Dr Richard M Royall, Department of Biostatistics, Johns Hopkins University, for helpful comments on the manuscript.

Funding: Swiss National Science Foundation (PROSPER 3233-32609.91).

Conflict of interest: None.

- 1 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;198:679-84.
- 2 Greenhalgh T. Statistics for the non-statistician. I. Different types of data need different statistical tests. *BMJ* 1997;315:364-6.
- 3 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.

- 4 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43-6.
- 5 Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 1995;142:904-8.
- 6 Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong RG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol* 1985;122:1080-95.
- 7 Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928;20A:175-240, 263-97.
- 8 Royall RM. *Statistical inference: a likelihood paradigm*. New York: Chapman and Hall, 1997.
- 9 Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc* 1942;37:325-35.
- 10 Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health* 1988;78:1568-74.
- 11 Oakes M. *Statistical inference*. Boston: Epidemiologic Resources, 1990.
- 12 Jones DR, Rushton L. Simultaneous inference in epidemiologic studies. *Int J Epidemiol* 1982;11:276-82.

(Accepted 16 January 1998)

## Coping with loss

### The doctor's losses: ideals versus realities

Glin Bennet

**This is the sixth in a series of 10 articles dealing with the different types of loss that doctors will meet in their practice**

44 Wellington Park,  
Bristol BS8 2UW  
Glin Bennet,  
formerly consultant  
psychiatrist, United  
Bristol Healthcare  
NHS Trust

Glin.Bennet@  
quadbtinternet.com

Series editors:  
Colin Murray  
Parkes and Andrew  
Markus

*BMJ* 1998;316:1238-40

After five years of study, newly qualified doctors may find it hard to realise that much of their future development will involve loss. They will go on gathering information and acquiring skills, but if they are to retain their enthusiasm and to mature as people, they will be learning to live with various losses.

#### Tiredness

New doctors should enjoy the initial enthusiasm, the ideals and the sense of omnipotence and invulnerability, the buoyant feeling of being able to contribute to the general good, because it may not last for long. Very likely a few months of broken nights will blur the ideals and push the ambitions into the distance. The immediate objective becomes to get through the job.

The grinding tiredness teaches them a lot: about their limitations, that sleep matters, and that it is difficult to be a good doctor when their eyes will not stay open. They become impatient over explanations, and tiredness comes up like a barrier so that they can no longer reach out to anxious and grieving patients.

They are learning that they cannot meet the ideals they set for themselves or the expectations of others. But tiredness is cured by a good sleep and enthusiasm is restored by a relaxing weekend. They can be admired for the long hours they work. They work harder than other people, they work amid the basic crises of living, they know about suffering, they see that people get better through their individual efforts, though they are not successful all the time. The death of a patient is a loss that reminds doctors of their limitations and the limitations of medical science, in which they had been taught to have so much faith. The first time it happens, the doctor is sad, shocked, perhaps angry that the patient could have done that to them.

#### Summary points

Reality often disappoints the expectations of young doctors, who become tired and disillusioned with themselves and with the health care system

A plateau in middle life is often associated with loss of further opportunities, and high achievers may interpret this as failure

To enjoy medicine we must achieve a balance between meeting the needs of our patients and maintaining our own resources of strength, energy, and commitment

Doctors who can acknowledge their own fallibility, accept their own wounds, and accept help from colleagues or others may emerge warmer and more humane

#### Loss of unreality

Most doctors have relatively simple lives in these early years, so it is possible, if they want, to give all their waking hours to the work in hand. Then there comes a time when the work is not sufficiently sustaining on its own—at least it ceases to be for most people, especially when the needs of others have to be considered. Now the people with the idealism and enthusiasm are confronted with a fresh reality, and much of a doctor's subsequent life and career will depend on how this matter is addressed.

This is a further lesson in the loss of omnipotence, but in no way is it the beginning of a decline. It is a time for redirecting energy. Doctors who accomplish this and can control the circumstances of their work can have a satisfying life, because medicine offers such abundant opportunities.



If, as doctors, we are to retain our enthusiasm and to mature as people, we must learn to live with loss

Many doctors make choices that put them in the front line, where they are directly exposed to needy members of the public. Here, the external pressures may seem always to be about to overwhelm them. The ideals are abandoned, and the redirection cannot be achieved. The giving out exceeds the individual's restorative powers. If as established doctors we find ourselves perpetually rushing to catch up with the demand as we see it, if our families tell us that we are irritable and our friends that we look tired all the time, then we are letting our resources become depleted: the signs of burnout will appear.

There is a phrase: "You have to be on fire before you can burn out." The idealism is gradually replaced by a mild cynicism, patients are perceived as inconsiderate and ungrateful, the telephone becomes an enemy. The process is familiar and, if unchecked, leads on to an apathy, in which minimum energy goes into the clinical work, although it may be redirected into administrative and extracurricular activities.

#### Men and women have different experiences

The traditional male approach has been to disregard these issues, at the outset of his career or later, and carry on up the professional ladder, regardless of personal considerations. If he becomes powerful enough he can sustain this style of life and put off his maturing, possibly forever. Sadly, these are qualities that make for professional success, and such people can be found among the more influential teachers of medical students, who are thereby exposed to the model of brashness and emotional immaturity.

Women doctors often experience loss before they qualify. They are affected more than most men by the brutalising aspect of medical education, which diminishes the empathic part of clinical work. They delay childbearing to their late 20s or beyond, and then they have to work that much harder to achieve the same professional goals as men.

#### Loss of meaning, loss of spirit

In the professional man's progress there are critical points, or hurdles, that can be cleared successively and that set him fair for the next stage. The loss of omnipotence is a necessary first stage. Another critical time for men comes in their late 30s, when they usually have their consultant post or their partnership in general practice. They are probably married with children, have a house, a good car, and a boat as well perhaps. The challenge here is that there are no more explicit challenges, such as qualifications and jobs; they have all been attained. Thus, there is a loss of these defining events and the excitement about all that is one day going to be. The doctor wakes one morning and says: "My life is now. This is what I am." This is the plateau of middle life.

High achievers imagine, albeit subconsciously, that the only way on from the plateau is downhill, and that once they have reached a particular professional peak there can only be loss of status, loss of role, diminishing health, and so on. The losses are real, but like all losses

they are points of transition, which can be seized creatively. There are gains from this new state: it is no longer necessary to meet all the expectations of others, to keep on achieving or publishing; it is no longer necessary to be wise and in control. There is more time. Life can broaden out at this point, provided that the losses are understood and seen for what they really are.

Women suffer loss in the process of becoming doctors, but there are great benefits later on, and unfulfilled male doctors could learn from their female colleagues. A woman doctor has professional work that she enjoys and often she has a family; thus she is not looking solely to her job to give her a life that is meaningful. Male doctors can do the same, only so often they avoid their families.

#### Doctors going wrong

When their life can no longer be defined by achievement, doctors often fail to cope. The work has lost its meaning because there is simply too much of it or because the doctor has not adapted to changing circumstances. The issues of omnipotence and invulnerability appear again, but in a more subtle form: it is no longer a matter of one not being able to accomplish everything, but rather accepting that one is merely a vulnerable human being. This comes hard to powerful doctors, and they may try to avoid the transition in various ways: by blotting it out (through alcohol); by reassuring themselves of their potency (with a new relationship); by channelling their energies in new directions (by getting on the train to London, where the important committees meet); by developing symptoms and becoming depressed.

All losses are points of transition, which can be seized creatively

Most problems that doctors experience in middle life are essentially problems of meaning, and their lives have lost their meaning because they have not been able to make the fundamental transitions and to value themselves simply for being who they are.



ADRIANA DORSETT

**Confidential advice and support**

There are currently many agencies for helping doctors. One example is the BMA's Stress Counselling Service (tel 0645 200169). It offers confidential telephone counselling 24 hours a day, 365 days a year, on a wide range of personal, emotional, and work related problems in areas such as:

- Debt and other financial concerns
- Workplace problems
- Difficulties with marital and personal relationships
- Alcohol and drug misuse
- Loss of confidence
- Stress and anxiety
- Bereavement.

Helpline numbers of two other agencies are 07071 223372 and 0121 558 0278.

losses are denied by them, so they are unable to make the essential transitions. It is as if they can accept themselves only as immaculate and all-competent professionals, and any blemish on that image is seen as a failure.

The losses are losses of illusions; the gains are gains in reality

**What is to be done?**

It is good that schemes have been organised whereby doctors in difficulties can get confidential help at a distance from their places of work, but there is a measure of failure here. Doctors become involved in such schemes because they have not been able to share their deeper feelings with their immediate colleagues, or because problems have been neglected until the situation is out of hand and the doctor is ceasing to function competently.

These doctors are stranded. They have of course lost their omnipotence and invulnerability, and they are no longer climbing the ladder to success, but these

In fact the blemishes have the potential to be a great advantage, and these doctors would do well to attend to the ancient idea that only the wounded physician heals. In myth it is presented in a literal form, but in ordinary life the "wounded" state refers merely to the acceptance of one's imperfections.

When doctors can accept their blemishes and vulnerabilities and their inability to achieve everything, they are free to make warm and ordinary relationships with their patients, family, and friends. They are free to look at the quality of their work and to make changes where these are needed. The losses are losses of illusions; the gains are gains in reality, and the quality of work and the quality of life can improve beyond recognition.

**Further reading**

Bennet G. *The wound and the doctor: healing, technology and power in modern medicine*. London: Secker and Warburg, 1987.

The articles in this series are adapted from *Coping with Loss*, edited by Colin Murray Parkes and Andrew Markus, which will be published in July.

### *A memorable patient* **Old remedy effective**

On a Shetland island, the home of hardy people, it was surprising to find that a runny nose and a cough meant that a visit to the doctor was required. And this was 40 years ago when life was less comfortable than today, when television was hardly known, and when people visited each other of an evening to talk and drink "a peerie glass o' the hame brew."

I was covering the period between the departure of the island's only doctor and the arrival of his successor. It was a wonderful six weeks among wonderful people. Their attitude to the common cold, though, was a bore. Mr X came to the surgery requesting a bottle of medicine for his cough. Explanation of the self limiting nature of his illness was treated with contempt. This being a dispensing practice I handed him some proprietary cough mixture from the big cupboard. No good. In two days he was back, seeking something more effective. My advice to hang on was again abruptly treated as nonsense and he went away with something of a different colour. Again, no good. It was too soon to consider investigating his cough, and as he was as determined as ever to take medicine I turned yet again to the cornucopia in the wall. There were now no more cough mixtures. There were, however, many big stock bottles of coloured glass containing ancient remedies with Latin names, and wedged among them, an out of date *National Formulary*. I sent Mr X home, promising to see if I could concoct something for him.

Mist Morph et Ipecac, with another now forgotten ingredient, was the result, the whole being dignified by the suffix "BPC." I made this up, using strange little measuring glasses graduated in

outdated units of measurement. I poured it into a bottle, stuck a cork in it, and labelled it with the instructions. It is with pride that I recall that I then took a dose of this concoction, believing that if I survived it he would. Mr X received this medicine and I heard no more till he returned four weeks later to have a loose tooth removed. This gave much less trouble than his cough but to the latter there was to be a sequel.

I took a phone call from an isolated cottage at the south end of the island. "Hello," a woman said, "I don't want you to come down but I've got a terrible cough and I just wondered if you would give the Postie a bottle of that medicine you made for Mr X to bring down to me. He says it's the same one that old Dr S used to make 30 years ago."

It is said that the only thing that distinguishes man from the lower animals is man's desire to take medicine. Especially it seems old medicines, and even in Shetland.

Bill Yule, *retired general practitioner, Forfar*

We welcome articles up to 600 words on topics such as *A memorable patient*, *A paper that changed my practice*, *My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.