*European Journal of Anaesthesiology* 1999, **16**, 73–76

EDITORIAL

# A case in point: exclusions and replacements for exclusions after randomization in clinical trials

In clinical trials, there may often be very good reasons why what is statistically desirable might not be clinically feasible: the sensibly pragmatic compromise of the clinical investigator becomes the statistician's nightmare. The paper by Gaitini *et al*. provides a useful focus for discussion of what to do when it becomes impossible or seems undesirable to complete a plan of investigation on every patient who has been randomly allocated to one group or another in a clinical trial, so that patients are excluded after the random allocation to groups.

In statistical parlance, Gaitini *et al*. set out to test the null hypothesis that, in healthy 1 to 8-year-old boys undergoing inguinal herniorraphy, supplementing the general anaesthesia with increments of intravenous fentanyl (control) affects the stress response to surgery no differently from supplementing it with a caudal block using local anaesthetic (test). For outcome variables, they measured changes in adrenaline and noradrenaline concentrations from baseline to indicate the intensity of the stress response during the operation and in the post-anaesthesia care unit. They decided on 20 patients in both the control and the treatment groups for reasons unspecified (although some sort of power calculation can reasonably be presumed).

Having allocated 20 patients randomly (details unspecified) into each group, they excluded four patients in the control group and three in the test group and replaced them to maintain the group sizes (without specifying how the new allocations were made). Three exclusions (all in the control group) were unavoidable because blood samples could not be obtained for one or more of the assessment times, so that the outcome variables could not be measured. One control patient was excluded because of hypothermia and two test patients because of an episode or more of hypoxaemia: these exclusion conditions were prespecified, presumably because their effects were thought likely to swamp any treatment effect on the catecholamine concentrations. One further patient was excluded because the caudal block did not produce effective intraoperative antinociception and could therefore be expected not to modulate the stress reponse.

The study and the analysis proceeded to give the results and conclusions published in the paper by Gaitoni *et al*. It appears, on the face of it, that plasma concentrations of adrenaline and noradrenaline tended to go up rather than down in the control group and down rather than up in the test group. A glance at Figs 1 to 4 might well tempt even a cautious gambler to place a substantial bet on it. The accompanying statistics indicated that the observed differences were very unlikely to have come about by chance under the hull hypothesis. Statistical analyses do not, as some may think, *prove* anything true or false. They attempt to obtain a precise estimate of the odds of being right or wrong in taking what is always going to be a gamble (namely whether or not to accept and/or act on a hypothesis).

If the odds for a null hypothesis are small enough in relation to other clinically important considerations, one may decide to reject it and believe that observed differences between two samples could not have arisen by chance. But this does not mean that we know how they *did* arise. Establishing causation is another matter from establishing 'significance'. It is the reason for randomization in clinical trials. Randomization is an attempt to remove all possible sources of known and unknown bias *except for* the treatment under consideration. This means that, if one decides to believe that differences do exist, they can reasonably be attributed to the treatment. Flaws in randomization do not call into question any conclusion about the statistical 'significance' of the observed differences, only the attribution to cause: flaws can allow known or unsuspected bias to enter the picture, which permits alternative attribution of cause.

Professional statisticians are almost unbelievably

© 1999 European Academy of Anaesthesiology


https://doi.org/10.1046/j.1365-2346.1999.00415.x Published online by Cambridge University Press

suspicious of bias. Bias is present until proved otherwise. It can insinuate itself through even the smallest flaw in a study design and, by so doing, contaminate conclusions about the cause of any differences for which chance is an unlikely explanation. The smaller the studies, the deeper the suspicions (and most clinical trials tend to be conducted with the smallest number of cases that will satisfy the power calculations for the effect size that is of interest). Subtracting cases after a randomized allocation is viewed as a substantial flaw, to be avoided at almost any cost. Adding cases to make up for exclusions after randomization threatens only to compound the problem (no matter how it is done). It should also be unnecessary. The best precaution against getting caught short by *unavoidable* losses of cases after randomization is to plan to have comfortably more cases in the study than the bare minimum required by the power calculations. The odd unavoidable exclusion would then still leave adequate numbers for power, and unequal losses between groups would rarely matter because it is not usually necessary to have exactly equal numbers in two groups that are to be compared in a clinical trial. Thus, Gaitini *et al*. would have done better to have settled on larger numbers and excluded without replacement the cases in which blood samples could not be obtained. But simply excluding them would not be enough: they would have to be examined characteristic by characteristic to check that they were representative of the cases that were included. The fear is that exclusion of unrepresentative cases might upset the balance of characteristics that has been created by the randomization. Outliers are difficult to detect confidently in small groups, which is another reason for generous initial allocations to groups.

*Avoidable* exclusions after randomization should not take place at almost any cost. The principle of 'intention to treat' should be sacrosanct. Cases should *all* be studied to the point of providing the prespecified outcome variable and should *all* then be included in the analysis. Prespecifying exclusions by criteria that can only come to light after randomization is an explicit contradiction of the principle of 'intention to treat'. The bias that they may admit may be more or less easy to identify or suspect. The reason for the exclusions by Gaitini *et al*. of the hypothermic and hypoxic cases is easy for a clinician to understand. However, this does not reassure a statistician that the attempt to remove one recognizable source of bias will necessarily remove exactly the right amount, and that it will not inadvertently introduce an unsuspected and unquantifiable additional source. In what should be rare exceptions to the principle of analysis by 'intention to treat', any exclusions for which outcome variables have or could have been obtained should be examined characteristic by characteristic, just as with the unavoidable exclusions, to check for inhomogeneities with the cases that have been included. This is especially important in trials that contain no more than the minimum numbers needed for the desired 'power'.

The preferred option is to base the study's definitive conclusions on the 'principal analysis' of data from all of the cases that have passed from randomization to provision of outcome variable. If there is concern about the inclusion of some cases (e.g. hypothermia, hypoxia, failed caudal block in the Gaitini paper), one should undertake a 'subsidiary' analysis to determine the 'sensitivity' of the definitive conclusions to the inclusion or exclusion of the contentious cases. If their exclusion supports the conclusions from the principal analysis at a similar level of confidence, so much the better: if the confidence is shaken to greater or lesser extent, this allows room for cautious qualification of the definitive conclusions. Further investigation of that qualification should provide the starting point for further study rather than being a finishing point of the current one.

Gaitini *et al*. would have been in a small minority of clinical investigators if they had followed these ideal recommendations to the letter. It is unlikely that they are in a position to do so now with the data that they have. Where does this leave them or, more importantly, their readers? There seems little doubt that the differences between their outcome variables at the various assessment times are highly unlikely to have arisen by chance under a null hypothesis. It would have been helpful if Gaitini *et al*. had quoted the results of their repeated measures analyses of variance (ANOVAS). Instead, they followed a disappointingly common practice of seeming to treat them solely as tiresome but necessarily preliminaries to the multiple testing for differences at several different time points. The repeated-measures ANOVAS would probably have shown a highly significant variance for the interaction term between treatment and time, addressing the important overall question of



https://doi.org/10.1046/j.1365-2346.1999.00415.x Published online by Cambridge University Press

whether the time course of the change in catecholamine concentrations in the test group was different from that in the control group. The multiple testing at individual times addresses the separate question about where in the time course any difference might be. It is perfectly possible to be very confident that two time courses are different without being very sure where exactly the differences are. The statisticians' doubts are about the effect of bias on the attribution of cause to the differences. Statisticians have the expertise to know how bias might *in general*, arise from exclusions and additions after randomization. Clinicians, and only clinicians, have the expertise to make a reasonable judgement about the likelihood of a *particular* source of bias in a *particular* clinical situation.

In a similar way, it is only the clinicians, not the statisticians, who can make a worthwhile judgement on how big an effect size is clinically important for the purposes of undertaking power calculations. Statistical and clinical expertise have equal weight in designing a clinical trials and in forming judgements on any difficulties that might arise. Clincians have a responsibility to contribute actively to a partnership with statisticians. Slavish acceptance of statistical dictats is as counterproductive as a stubborn refusal to understand the intent of statistical advice.

In considering the possible causes, other than the treatment effect (caudal v. fentanyl), for the differences observed by Gaitini *et al.*, we clinicians should try to consider the likelihood that each particular exclusion might have introduced a bias. We should consider the specific nature of the possible bias and the likelihood that it could provide an alternative explanation for, or contribute to or subtract from the observed differences. For example, how likely is it that the following statements are true?

(a) Blood samples are more difficult to obtain under anaesthesia with fentanyl supplementation, and the three patients who were excluded from the control group for this reason would independently have had lower catecholamine concentrations than the control patients whose results were analysed.
(b) Patients are more likely to become hypothermic with fentanyl supplementation, and the patient who was excluded from the control group for this reason would independently have had lower catecholamine concentrations than the remainder of the control group.
(c) Patients are more likely to become hypoxic with caudal blockade, and the two patients who were excluded from the test group for this reason would independently have had higher concentrations of catecholamines than the other test patients.
(d) Patients in whom it is difficult to establish successful caudal anaesthesia (and who would not have been excluded from the control group) tend, independently of the ineffectiveness of the caudal block, to have higher catecholamine concentrations than do patients in whom effective caudal block can be established.

If, after giving these sorts of statements their careful expert consideration, clinicians conclude that the statements are unlikely to be true, they might decide to gamble on believing that the treatment (caudal v. fentanyl) *was* responsible for the observed alterations in physiology. Even if the study design and associated statistics had been impeccable, it would still have been a gamble: the odds may or may not have been different, but they would probably have been more precisely defined. In deciding whether or not to change their practice after accepting a difference in physiology, clincians would have to embark on an entirely separate gamble on whether that difference is associated with a sufficiently important change in clinical outcome for the patient. The study by Gaitini *et al.* is not intended to offer any help with this decision.

## Conclusions

*A peppering of pin-head P-values positively prognosticates probable publication – BUT*

(1) Even impeccable statistics based on impeccable study design do not *prove* anything. They simply ensure precision in estimating the odds of being wrong in believing and acting.
(2) Randomization in clinical trials is done to minimize the risk of bias in attributing cause to differences that seem unlikely to have arisen by chance under a null hypothesis.
(3) Exclusions after randomization in clinical trials are a potential source of bias.
(4) All avoidable exclusions after randomization *should* be avoided at almost any cost, all cases



https://doi.org/10.1046/j.1365-2346.1999.00415.x Published online by Cambridge University Press

should be followed through to the outcome and the definitive conclusions from the study should be based on analysis of outcome by 'intention to treat'.

(5) Replacements for unavoidable exclusions should be avoided. Power calculations should be regarded as determining only the *minimum* group sizes for adequate resolution of a clinically important effect size: an allowance for unavoidable exclusions should be added.

(6) As many as possible of the characteristics of excluded cases should be compared with those of the cases that are included. Additional generosity with group sizes is needed for one to be at all confident in recognizing that exclusions may be outliers.

(7) A subsidiary 'sensitivity' analysis should be undertaken to investigate the effects of excluding cases for which exclusion can reasonably be argued. This may qualify the definitive conclusions on causation, but such qualifications should be the tentative preliminaries to further study.

(8) Statistical considerations have no sole right to dictate planning or publication of clinical trials. A partnership of statistical and clinical expertise is vital. Studies whose design is irretrievably flawed are not necessarily worthless. Individual clinicians should use their *specific* expertise and experience to come to a judgement on whether statisticians' *general* suspicions of bias are justified in the *particular* circumstances of the clinical trial being considered.



https://doi.org/10.1046/j.1365-2346.1999.00415.x Published online by Cambridge University Press