Pergamon

European Journal of Cancer, Vol. 33, Suppl. 2, pp. S14–S16, 1997

Printed in Great Britain
0959–8049/97 $17.00+0.00

PII: S0959-8049(96)00449-2

A Plea for Improved Use of the Tools Available for the Evaluation of Anticancer Drugs

K. MacRae

Charing Cross and Westminster Medical School, The Reynolds Building, St Dunstan's Road, London W6 8RP, U.K.

Superb methods of anticancer drug evaluation exist, but they are still being used inadequately. In comparative trials, it is essential that the statistical design and analysis eliminate the possibility of allocation (group membership) bias, assessment (measurement) bias and chance before it is possible to conclude that a real difference exists between the two treatments under test. Oncologists should examine these aspects critically before accepting that a trial shows a genuine treatment effect. © **1997 Elsevier Science Ltd. All rights reserved.**

Key words: bias, chance, power, randomisation
Eur J Cancer, Vol. 33, Suppl. 2, pp. S14–S16, 1997

INTRODUCTION

SUPERB METHODS are available for evaluating the efficacy of cancer treatments. Problems arise, however, because oncologists carrying out trials may either use the wrong methods or badly use the right methods.

In my view, a factor contributing to this is the (il)logical construction derived from the hypothesis that if the treatment is effective, the patient will recover: if the patient does recover, therefore, it is argued that the treatment must have been effective. This is known in logic as the fallacy of affirming the consequent, but in medicine it is often termed 'clinical experience'.

A corollary of this is that the open, uncontrolled study may yield highly misleading results. A striking example of this is provided by review of a series of 20 uncontrolled studies of the same treatment (rapid injection of 5-fluorouracil (5-FU)) in the same disease (advanced cancer of the colon) using the same outcome measure ('objective regression'), but yielding success rates ranging anywhere between 8 and 85% [1].

The method used in oncology to overcome this difficulty is the comparative clinical trial, often called a randomised, controlled trial. Used properly, this is extremely powerful, but the potential problems can all be illustrated by considering the set of results in Table 1. Despite the apparently large difference in success rate between the two treatment groups, it is only possible to conclude that this difference is a real one when the three other possible explanations of allocation or group membership bias, assessment or measurement bias and chance have been eliminated.

ALLOCATION (GROUP MEMBERSHIP) BIAS

A difference between treatment groups sometimes arises because the patients, rather than the treatments, are different. This is a substantial problem in oncology today.

Allocation bias may be eliminated from the trial design by ensuring that determination of patient eligibility for the trial is carried out blindly and by using one of several possible bias-free methods to assign patients to treatment groups. Of the methods available, random allocation is most often used, but others include stratified randomisation, minimisation and cross-over designs.

The elimination of allocation bias from the original design of trials is, therefore, relatively straightforward, but another, more subtle source of bias is that created by the exclusion of certain patients from the final analysis. By taking patients out of the analysis, the bias is put back into the comparison. For this reason, trials should always include an intention-to-treat analysis.

In practice, however, the intention-to-treat analysis is only one of three analysis policies widely used. Instead, investigators often analyse on an all patients treated (APT) or per protocol (PP) basis. In an APT analysis, only those patients who received some treatment are included—those who were randomised but received no treatment are left out. When results are analysed PP, only the 'evaluable' patients (i.e. those patients who met all the protocol requirements and received all the treatments planned) are included.

Both these alternative policies have pitfalls, which are highlighted by the hypothetical results in Table 2. In this trial, climbing a mountain was compared with the standard combination of cyclophosphamide/methotrexate/5-FU (CMF) as a treatment for cancer. The results with CMF were straightforward, with 57 out of 100 patients (57%) still alive after 12 months. In contrast, mountain-climbing was 100% successful.

Table 1. A hypothetical set of results

	Active drug	Placebo
Success	19	3
Failure	1	17
Total	20	20

Table 2. Mountain-climbing versus cyclophosphamide/methotrexate/5-FU (CMF) as a treatment for cancer

	CMF	Mountain-climbing
n	100	100
12-month survival (%)	57	100

Table 3. Breakdown of results for mountain-climbing as a treatment for cancer. From the table, 12-month survival is 9/9 (100%)

n	100
Refused consent	28
Lost to follow-up	34
Non-compliers	29
Evaluable patients	9

The results in the mountain-climbing arm, however, were complicated to analyse. Of 100 patients randomised, 28 refused to consent and did not begin the climb, 34 became lost and did not complete follow-up, and another 29 non-compliers were found either half-way up or half-way down the mountain and had not, therefore, completed the full course of treatment. Ultimately, therefore, only 9 patients fulfilled all the protocol requirements, and they were all alive 12 months later (Table 3).

These illustrations demonstrate the problems arising from patient exclusions and highlight the particular need for caution in accepting a conclusion that is often drawn in oncology studies. This conclusion is that patients who take the full course of treatment, or complete the full cycle of treatment, tend to fare better than those who do not. The danger inherent in the assumption that treatments only work if patients comply with them has been previously identified outside the field of oncology. With the lipid-lowering agent clofibrate, for example, analysis by compliance revealed that mortality among patients who complied with more than 80% of the prescribed dose was greatly reduced relative to that in non-compliers, but the same was also true for placebo [2]. Should we interpret this as meaning that taking a placebo daily will reduce our risk of death?

Clearly not. Rather, as the authors point out, these findings illustrate forcefully the point that compliance and non-compliance were not determined by random allocation (i.e. group membership bias was present) [2]. The patients made a choice as to whether or not they took the tablets, whichever type they were (active or placebo). They probably also made related choices about taking other steps that might reduce their mortality risk, such as stopping smoking, drinking less alcohol or taking more exercise.

MEASUREMENT BIAS

Clinicians often focus attention on the nature of the outcome measure, and it is important that this is both valid and relevant. The need for the outcome to be free from measurement bias, however, is sometimes overlooked. This is the bias that arises when an outcome, such as 'response' or 'success', is judged by different people to different standards. Outcomes must, therefore, be capable of objective measurement (and this is a strong argument in favour of the use of survival as an endpoint) or, if measures such as clinician judgement or patient opinion are used, then the assessors of these outcomes must be blinded. To date, blind studies in medical oncology have been quite unusual, yet even surgeons may conduct double-blind trials by using independent assessors who are not informed which operation was carried out.

CHANCE

To eliminate chance as an explanation of a study's results, it is essential that the study is designed with a size appropriate to answer the question being asked *and* that the *P*-value calculated in the statistical analysis of the data is sufficiently low for the results to be considered significant. Herein lie many potential pitfalls, presenting major problems in oncology at present. Currently, study reports often include interim analyses, subgroup analyses or multiple comparisons that were not planned in the original design. Subgroup analyses are, perhaps, the most dangerous of these: if four separate subgroups are analysed, the likelihood that at least one statistically significant result will be obtained is 18.5%. If 10 subgroups were analysed, there would be a 40% chance of a significant result [3]. The old statistical saying—if you torture the data long enough it will confess—highlights the danger that investigators are tempted to manipulate the data until it yields a positive result.

P-*value*

Poor understanding of the meaning of the *P*-value is sometimes a source of certain difficulty in interpretation. A *P*-value is the probability of obtaining the data by chance, given the hypothesis that the treatments are the same, *not* the probability that the treatments are the same, given the data. Given the choice between the correct definition of *P* and the fallacy of the transposed conditional, most non-statisticians choose the fallacy.

The *P*-value is calculated using a statistical test, such as the chi-squared test. A small difference between treatments will yield a relatively high *P*-value, while, for a large difference, the *P*-value will be lower. This reflects the fact that when two treatments are the same, a small difference may quite often be seen between the patient groups, but only occasionally a large difference. There is a convention, which we owe to Professor Sir Ronald Fisher FRS, that $P < 0.05$ means that the data are statistically significant and we may reject chance as an explanation.

Study size

Table 4 reviews the seven major studies comparing treatment of early breast cancer using surgery plus radiotherapy with surgery alone. It clearly illustrates the problem in respect of the

Table 4. The seven major studies comparing treatment of early breast cancer using surgery plus radiotherapy with surgery alone

Study [Ref.]	Total	Lower % success/survival	Upper % success/survival
CRC [4]	2268	50	57.5
Manchester P [5]	741	50	64
Manchester Q [5]	720	50	64
Edinburgh [6]	385	50	68
Guy's [7]	370	50	68.5
Cambridge [8]	204	50	74
Copenhagen [9, 10]	400	50	68

This table shows the improvement from an arbitrary 50% success or survival rate that each study could have detected at the 0.05 level of significance and a power of 95%.

large study sizes necessary in cancer trials. The Cancer Research Campaign (CRC) Trial, which recruited over 2000 patients, was designed to have a greater than 90% chance of detecting a 7% difference in survival between the two treatment groups [4]. The other studies were only capable of detecting differences that were unrealistically large [5–10]. The Cambridge trial, for example, examined only 200 patients, which gave it power to detect a 24% difference in absolute success rates [8]. Expecting such a large difference between two different forms of local treatment for early breast cancer is, at least, optimistic.

CONCLUSION

When used correctly, methods of drug evaluation based on comparative clinical trials provide a strenuous test of the effectiveness of a given drug of treatment for cancer. Considerable care must be devoted, however, to the statistical design and analysis of the trial if the results are to have genuine rigour. For convenience, the issue of whether or not the difference between treatments is real may be summarised in the form of five separate questions:

- Has the design prevented allocation or group membership bias (i.e. is the trial randomised, or does it use, for example, 'historical' controls)?
- Has the design prevented assessment or measurement bias (i.e. are the data hard outcomes or have non-blinded and biased assessments been used)?
- Has the analysis prevented group membership bias (i.e. has bias been reintroduced by leaving some patients out)?
- Has the analysis made a chance effect more probable than the level of significance (i.e. have the data been 'tortured')?
- Is the *P*-value significant?

A final word of warning: meta-analysis, the pooling of the results of all trials directed at answering the same question, is a useful development, but one that needs extremely cautious interpretation. It is outside the scope of this review.

1. Moertel CG, Reitemeier RJ. *Advanced Gastrointestinal Cancer: Clinical Management and Chemotherapy*. New York, Harper and Row, 1969.
2. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *New Engl J Med* 1980, **303**, 1038–1041.
3. MacRae K. Pitfalls in interpreting the results of cancer trials. *Cancer Care* 1987, **4**, 4–7.
4. Cancer Research Campaign Working Party, Cancer Research Campaign (King's/Cambridge) trial for early breast cancer. A detailed update at the tenth year. *Lancet* 1980, **2**, 55–60.
5. Easson EC. Post-operative radiotherapy in breast cancer. In Forrest APM, Kunkler PB, eds. *Prognostic Factors in Breast Cancer*. Edinburgh, Livingstone, 1968, 118–127.
6. Bruce J. Operable cancer of the breast: a controlled clinical trial. *Cancer* 1971, **28**, 1443–1452.
7. Atkins H, Hayward JL, Klugman DJ, Wayte AB. Treatment of breast cancer: a report after ten years of a clinical trial. *Br Med J* 1972, **2** 423–429.
8. Brinkley D, Haybittle JL. Treatment of stage-II carcinoma of the breast. *Lancet* 1966, **2**, 291–295.
9. Kaae S, Johansen H. Breast cancer: five year results; two random series of simple mastectomy versus extended radical mastectomy. *Am J Roentgenol, Rad Therap & Nuclear Med* 1962, **87**, 82–88.
10. Kaae S, Johansen H. Simple versus radical mastectomy in primary breast cancer. In Forrest PM, Kunkler PB, eds. *Prognostic Factors in Breast Cancer*. Edinburgh, Livingstone, 1968, 93–102.