Acta Oncologica 26 (1987) Fasc. 6

FROM THE DEPARTMENT OF ONCOLOGY AND RADIOTHERAPY, ODENSE UNIVERSITY HOSPITAL, DK-5000 ODENSE, DENMARK.

PROBLEMS ASSOCIATED WITH COMPARISONS OF RESPONSE-DEFINED SUBSETS OF PATIENTS IN RANDOMIZED TRIALS

Treatment-related bias and response migration

H. BRINCKER

Abstract

Subset analysis may be justified between various arms of randomized trials as long as subsets are defined by variables which do not cause distortion of other prognostic factors. However, bias will occur when treatment response is used to define a subset of patients in which the results of the same treatment are analyzed. Thus, comparisons between groups of responders in randomized studies are just as inherently biased as comparisons between responders and non-responders. Using a constructed example the effects of treatment-related bias on the interpretation of trial results are demonstrated, and the concept of response migration is introduced. It is shown that in randomized trials the only unbiased measures of treatment efficacy are response rate, overall time to progression, and overall survival.

Key words: Randomized trials, response-defined subset, bias.

It has rightly been emphasized that survival comparisons between responders and non-responders often presented in reports on non-randomized chemotherapy trials are biased and may lead to misleading conclusions regarding treatment efficacy (1–3, 5–9). So far, however, the validity of survival comparisons between patients belonging to the same response category but to various arms of a randomized study has not been questioned. Indeed, one of the papers criticizing the comparison of responders and non-responders states that 'reports of phase III trials ... may compare responders in the different groups' (9). The general consensus appears to be in support of this statement, and many randomized studies present survival comparisons between responding patients usually with statistical tests. It often seems to be implied that a 'statistically significant' difference supports a cause-effect relationship between treatment and survival. Nevertheless, such comparisons may actually be biased since treatment response itself is used to define a subset of patients in which the results of the same treatment are analyzed. The basis of a valid comparison between two differently treated groups in a randomized study is the assumption that patients were randomly assigned to the groups with treatment being the only difference. This condition is certainly fulfilled when all patients randomized to one group are compared with all patients randomized to another group. However, when we compare response-defined subsets of patients, be it responders or non-responders, the distribution of prognostic factors is almost certainly not the same in the groups that are being compared due to different mechanisms of action of the two treatments, due to different sizes of the groups resulting from different response rates, or due to other treatment-related factors. To complicate matters further, conventional survival curves give a false impression of the true difference between response-defined subsets when the response rates are different in these subsets. The reason is that when different curves all start at 100% on the y axis of the survival graphs, it is wrongly implied that the areas under the curves are directly comparable.

In order to illustrate these problems of treatment-related bias and of interpretation of survival curves in response-defined subsets in randomized trials a constructed example will be analyzed. The example intends to show how treatments of different effectivity may affect comparisons between patients belonging to the same response category, but to various arms of a randomized study.

An example. For the sake of simplicity we will assume a

Accepted for publication 22 July 1987.

study of 120 patients, randomized into 4 groups of 30 patients each, receiving treatments A, B, C and D respectively (Table 1). In group A we observe 18 responders (=complete plus partial remission) and 12 non-responders (=no change plus progressive disease). In groups B, C and D we observe 12 responders and 18 non-responders. Thus, the response rate is 60% in group A versus 40% in groups B, C and D. We assume, further, that progression-free survival times of the patients in group A will vary from 1 to 30 months. The same applies to group B. Thus, the latter group has 20% fewer responses than group A, but it does not affect the overall median time to progression (TTP). In group C we will assume an overall median TTP of 3 months (or 19.4%) less than in group A. Similarly, the overall median TTP in group D will be 6 months (or 38.7%) less than in group A. In other words, A is the most 'effective' treatment, and the effectiveness of the treatments decreases progressively from A through D. With groups of only 30 patients the differences between groups B and C and C and D are not statistically significant ($p=0.14$). However, the differences could easily be made statistically significant just by increasing the size of the groups.

For didactic reasons TTPs have been chosen so that they progress in steps of 1 month, from 1 month in non-responders to 24 to 30 months in responders. As a result of this artifice all survival curves will be simplified considerably, appearing as straight lines. 'True' survival curves will usually have a 'tail', but no attempt has been made to reproduce true curves, since this is irrelevant for the discussion.

All four treatment groups in Table 1 have arbitrarily been divided into 3 subgroups of 10 patients each (by the dotted lines). Response to treatment decreases progressively from subgroup I through subgroup III. Factors predicting response to treatment include known variables, such as extent of disease, histology, performance score, etc., as well as unknown variables. The composition of the 3 subgroups is identical in the 4 treatment groups, A–D, since the patients are assumed to be randomized. Each of the 3 subgroups will contain patients in whom the response to treatment is predicted to be good (index 3), intermediate (index 2), or poor (index 1). An index of 3 or 2 indicates that a good or moderate response is predicted, but not that it will invariably occur. Conversely, the prediction of a poor response (index 1) does not preclude an occasional good response. Thus, for treatment A index 3 correctly identifies 8 out of 10 expected responders, index 2 identifies 7 out of 10 expected responders, and index 1 identifies 7 out of 10 expected non-responders. For treatments B through D the corresponding figures are 7, 3 and 8 respectively. It is, of course, assumed that the number of patients with response to treatment decreases from subgroup I through subgroup III. The mean index predicting response is 2.3 for subgroup I, 2.0 for subgroup II, and 1.7 for subgroup III.

Table 1

*Progression-free survival (months) and type of response in 120 patients, randomized to 4 different treatments. *Index 3 = predicted good response, index 2 = predicted intermediate response and index 1 = predicted poor response*

Patient No.	Treatment				Prognostic subgroup	Index predicting response*
	A	B	C	D		
1	30	30	27	24		3
2	29	29	26	23		3
3	28	28	25	22		3
4	27	27	24	21		3
5	26	26	23	20	I	3
6	25	25	22	19		2
7	24	24	21	18		2
8	23 responders	23	20	17		2
9	22	22	19	16		1
10	21	21	18	15		1
11	20	20	17	14		3
12	19	19	16	13		3
13	18	18	15	12		3
14	17	17	14	11		2
15	16	16	13	10	II	2
16	15	15	12	9		2
17	14	14	11	8		2
18	13	13	10	7		1
19	12	12	9	6		1
20	11	11	8	5		1
21	10	10	7	4		3
22	9 non-	9	6	3		3
23	8 responders	8	5	2		2
24	7	7	4	1		2
25	6	6	3	1	III	2
26	5	5	2	1		1
27	4	4	1	1		1
28	3	3	1	1		1
29	2	2	1	1		1
30	1	1	1	1		1

Some consequences. Fig. 1 shows a conventional comparison between treatments A and B. Overall median TTP is, of course, the same in both arms. However, for both responders and non-responders median TTP of B patients is 3 months longer, and if we disregard the comparison of overall TTP we might be tempted to conclude that treatment B is better than treatment A, although the latter resulted in 20% more remissions. The reason for this apparent paradox is, clearly, that treatment A causes some patients, who would have been non-responders with short TTPs on treatment B, to migrate into the group of responders, thereby reducing the median TTP for both responders and non-responders.

In Fig. 2 treatments A and C are compared. As planned, overall median TTP is 3 months (or 19.4%) longer for A patients than for C patients, documenting a slight superiority of treatment A. Nevertheless, if we look at responders and non-responders separately, median TTP is identi-

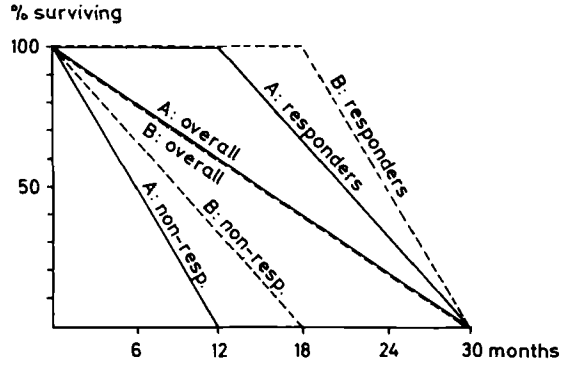

Fig. 1. Treatment A versus treatment B. Overall time to progression, and time to progression for responders and non-responders. Abbreviations: % surviving = % surviving without progression. Non-resp. = non-responders.

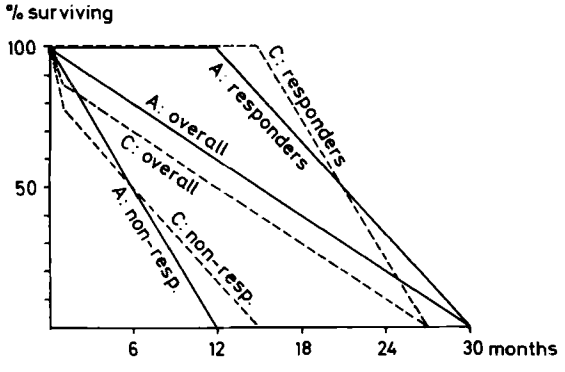

Fig. 2. Treatment A versus treatment C. See text to Fig. 1.

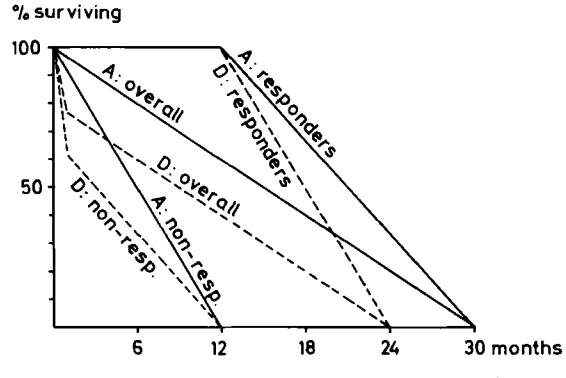

Fig. 3. Treatment A versus treatment D. See text to Fig. 1.

cal for treatments A and C. We might therefore be misled to conclude that treatments A and C are equally effective, unless we also compare overall TTP.

Fig. 3. compares treatments A and D. These two treatments are so different that median TTP is longer for treatment A than for treatment D, regardless of whether we compare responders, non-responders, or all patients. However, it is noteworthy that the difference between

Table 2

Mean index predicting response in various subgroups of patients

	Mean index	
Subgroup		
I	2.3	All patients = 2.0
II	2.0	
III	1.7	
A (responders)	2.28	
B, C, D (responders)	2.42	
A (non-responders)	1.58	
B, C, D (non-responders)	1.72	

median TTP in treatments A and D is only 3 months in responders versus 6 months in all patients.

Comparisons between groups of patients with the aim of measuring the relative efficacy of various treatments assume that the groups that are compared are truly comparable. Accordingly, if we compare subgroups of patients, such as responders, between two arms of a randomized study, the comparison is only unbiased if we have the same distribution of factors predicting response in the two groups. Table 2 compares the mean response indices of responders and non-responders between group A and groups B, C and D. Neither in responders, nor in non-responders the same distribution of factors predicting response is found. As might be expected, the greatest difference is found between responders on one side and non-responders on the other.

Discussion

The results summarized in Figs 1–3 and in Table 2 show that a conventional comparison of median TTP (or median duration of response) between responders in various arms of randomized studies a) may lead to erroneous conclusions regarding relative treatment efficacy, and b) suffers from essentially the same type of bias as comparisons between responders and non-responders. Clearly, median TTP of responding patients tells us how long a response will last with a particular treatment if a response is obtained, but it cannot be viewed outside the context of the response rate that was obtained. In fact, the apparent paradox of the 'best' treatment having a median TTP for responders which may be shorter (Fig. 1), identical with (Fig. 2), or longer (Fig. 3) than that of the 'worst' treatment is due to response migration, analogous to stage migration, recently named 'the Will Rogers phenomenon' (4).

The constructed example might be easier to visualize if we consider the expected results of conventional versus intensive therapy for advanced breast cancer. If oophorectomy only is used the expected response rate is about 30%, but with combination chemotherapy such as CAF

(cyclophosphamide/adriamycin/5-fluorouracil) the response rate may be 60%. About 50% of oophorectomy responding patients remain in remission after 1 year, compared with only about 30% of CAF patients. In this situation the more aggressive CAF therapy converts poorer prognosis patients, who would not have responded to oophorectomy, into response without having a substantial impact on their overall survival. Since the CAF responders are a mix of patients whose prognosis is poorer than that of the oophorectomy responders, a comparison of the TTP of responders could lead to the incorrect conclusion that CAF is a poorer therapy.

Subset analysis may be justified between various arms of randomized studies as long as subsets are defined by variables that do not cause distortion of other prognostic factors. For example, if we analyze response rates according to stage of disease, it is likely (if the study is large enough) that within each stage there will be an identical distribution of histologic types, performance scores, etc. However, if we use treatment response itself to define a subset in which we analyze the results of the same treatment, there is no longer any guarantee that factors predicting response or survival will be distributed in the same way, neither in responders nor in other response-defined subsets in various arms of the study. This is clearly demonstrated in Table 2. It has now become widely accepted that comparisons between responders and non-responders are biased (1–3, 5–9), but the case against comparing responders with responders in randomized studies is also strong. In fact, it follows from Table 2 that the comparisons between responders and non-responders respectively, shown in Figs 1–3 must also be subjected to this criticism since these groups are not truly comparable.

Strictly speaking, only the *entire* groups originally randomized can be compared. This means that in randomized trials there are only three completely unbiased measures of treatment efficacy, namely response rate, overall TTP and overall survival. However, the latter parameter may be compromised by subsequent treatment, for example if the study contains a cross-over design. Clearly, the bias that results from comparing 2 groups of responders is likely to be less than that resulting from comparing responders with non-responders (Table 2). Nevertheless, the problem remains that the size of the bias cannot be accurately assessed. Accordingly, comparisons between groups of responders give only an indication, not a direct measure of actual treatment differences. Furthermore, if a statistical test shows a significant difference between two (or more) groups of responders, this cannot be used to support a cause-effect relationship between treatment and response duration. Figs 1–3 demonstrate that overall TTP is a more accurate indicator of true treatment differences than TTP of responders. Since overall TTP is also an unbiased parameter it should be used routinely instead of TTP of responders.

In conclusion, comparisons of any response-defined subsets of patients in randomized trials are inherently biased and should be avoided, or at least be interpreted with great caution. In addition, conventional survival curves may give a false impression of the true differences between response-defined subsets due to the phenomenon of response migration.

Request for reprints: Dr Hans Brincker, Department of Oncology and Radiotherapy, Odense University Hospital, DK-5000 Odense, Denmark.

REFERENCES

1. Aisner J. and Hansen H. H.: Commentary. Current status of chemotherapy for non-small cell lung cancer. Cancer Treat. Rep. 65 (1981), 979.
2. Anderson J. R., Cain K. C. and Gelber R. D.: Analysis of survival by tumor response. J. Clin. Oncol. 1 (1983), 710.
3. — and Davis R. B.: Letter to the Editor. J. Clin. Oncol. 4 (1986), 115.
4. Feinstein A. R., Sosin D. M. and Wells C. K.: The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. New Engl. J. Med. 312 (1985), 1604.
5. Mantel N.: An uncontrolled clinical trial—Treatment response or spontaneous improvement? Controlled Clin. Trials 3 (1982), 369.
6. Oye R. K. and Shapiro M. F.: Reporting results from chemotherapy trials. Does response make a difference in patient survival? J. Amer. Med. Ass. 252 (1984), 2722.
7. Schneiderman M. A.: Non-objective and objective evaluation in cancer chemotherapy. *In:* Cancer Chemotherapy. Basic and Clinical Application, p. 67. Edited by I. Brodsky and S. B. Kahn. Grune and Stratton, New York 1967.
8. Simon R. and Wittes R. E.: Methodologic guidelines for reports of clinical trials. Cancer Treat. Rep. 69 (1985), 1.
9. Weiss G. B., Bunce H. and Hokanson J. A.: Comparing survival of responders and non-responders after treatment. A potential source of confusion in interpreting cancer clinical trials. Controlled Clin. Trials 4 (1983), 43.