Analysis of Survival by Tumor Response

By James R. Anderson, Kevin C. Cain, and Richard D. Gelber

The common practice of comparing the survival of responders and nonresponders when reporting the results of cancer chemotherapy treatment is investigated. The usual method of comparing responders and nonresponders is biased in favor of responders, and these results are frequently misinterpreted as providing evidence that response prolongs survival, or that the treatment under study is effective. Two valid methods for comparing responders and nonresponders are discussed and recommendations are made concerning the analysis of survival by response. A comparison of survival by response category may be useful descriptively, but such a comparison should not be used for inference concerning treatment effectiveness.

IT IS a common practice for investigators to present comparisons of survival for responders versus nonresponders when reporting the results of chemotherapy regimens for cancer patients with advanced measurable disease. In many papers, these results represent the major ''statistically significant'' finding of the study. Investigators often conclude that if responders survive significantly longer than nonresponders, then the effect of response is to prolong survival. They often further conclude that, when response ''prolongs'' survival, increasing the response rate will be a way to increase survival.

Many of the papers that compare survival for responders versus nonresponders contain two serious errors. The first error is in the statistical method used for detecting a significant difference between the survival of the two groups. The standard methods are biased and can produce a ''significant'' difference when none in fact exists. The reasons for this are discussed later, and two correct methods are described.

The second common error is that the demonstration of a significantly longer survival for responders is frequently misinterpreted to imply that response causes longer survival. This conclusion might seem quite reasonable biologically, but the data do not necessarily imply causation. It is generally impossible to refute the possibility that response is just a marker which selects the good prognosis patients: those who would have survived longer even if the therapy has no effect at all. A further misinterpretation often implicit in survival comparisons is that longer survival for the responders implies that the treatment is effective.

Finally, recommendations for the analysis of survival by response are made. The main conclusion is that the comparison of responders versus nonresponders can be quite often useful descriptively (for clinical practice), but such a comparison should not be used for inferences about treatment effectiveness.

THE ANALYSIS OF RESPONSE AND SURVIVAL

The effectiveness of chemotherapeutic regimens for patients with advanced cancers is assessed using several endpoint measures. The two most frequently used are objective tumor response and overall survival from start of treatment. Objective tumor response is often categorized by one of four states: complete response, partial response, no response, and progressive disease. Occasionally, the first two categories are combined to define responders and the last two combined to define nonresponders.

Survival by tumor response category is often a part of an analysis of clinical trial data. Patients are categorized as either responders or nonresponders, and the life-table estimated survival for each group is calculated, usually measured from the start of treatment. These survival curves are then compared, and the ''statistical signifi-

From the Department of Biostatistics and Epidemiology, Dana-Farber Cancer Institute, and the Department of Biostatistics, Harvard School of Public Health, Boston, Mass.

Submitted March 30, 1983; accepted July 22, 1983.

Supported in part by grants no. CA-19589, CA-23415, CA-09337, CA-33601, and CA-23318 from the National Cancer Institute, National Institutes of Health, Bethesda, Md.

Address reprint requests to James R. Anderson, PhD, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.

0732-183X/83/0111-0008$2.00/0

cance'' of the observed differences is determined by a log rank test, or another appropriate significance test.

The frequency of published survival comparisons by response category was investigated by a review of the first nine issues of *Cancer* published in 1982. Thirty-one reports described the treatment of patients with advanced cancers. Fifteen of these reports contained some details of a comparison of survival for responders versus nonresponders. In nine of these reports (29%), the *p* value for the comparison was presented as a result in the analysis. Thus, the analysis of survival by response category is quite common in the published literature.

STATISTICAL METHODS FOR DETECTING A DIFFERENCE IN SURVIVAL BETWEEN RESPONDERS AND NONRESPONDERS

The Usual Method

The most frequently used method for analyzing survival by response involves separating patients into two groups according to whether or not they ever achieve a response. Life-table estimates of the survivorship function (survival curve) are computed for each group, and a statistical test such as the log rank test[1] is used to determine if the difference between the survivorship functions is statistically significant.

A bias exists in which the length of survival itself will influence the chance of a patient being classified into one group or the other. Patients who eventually become responders must survive long enough to be evaluated as responders. This so-called guarantee time is at least as long as the time to the first response evaluation. This requirement of a longer survival time also provides a greater opportunity for the therapy to produce a response. No such guarantee time is required for the nonresponder group. In fact, patients who die during the period before the first response evaluation are automatically included in the nonresponse group. Thus, patients with poor survival prognosis who die early in the study will not have an opportunity to enter the responder group and will guarantee poorer survival for the nonresponse group. As an extreme example of this bias, suppose that the first response evaluation is done at two months from on-study time, and that virtually all of the patients still alive at two months have achieved a response. Responders are guaranteed at least a two-month survival, and patients who die within two months are automatically labeled nonresponders. The conclusion that responders live longer than nonresponders is the same as saying that those patients who live at least two months survive longer than those who do not. This is undeniably true, but not very interesting.

This bias caused by the guarantee time for responders directly affects the validity of statistical tests used to compare the survivorship functions, in that responders are considered to be at risk of failure during the entire follow-up period. In fact, patients who eventually respond are at risk of failure from the responder category only after their status as responders is determined. Counting patients who are responders as being at risk of failure before the time of response will result in an inappropriately favorable survival curve for responders, and an inappropriately unfavorable survival curve for nonresponders. As statistical tests such as the log rank test assess the equality of survivorship functions by comparing the estimated death rates over time,[1] the usual method is biased in favor of the responders. To illustrate this phenomenon, we consider one week's contribution to the log rank test in a hypothetical example (Table 1). Assume that at the start of the seventh week of treatment, there are 100 patients still alive, and that 50 of these patients either have already achieved a response or will achieve a response at some time in the future. Suppose that 10 patients die during the seventh week, of which two are responders and eight will never respond. It thus appears that the death rates for week 7 are 4% for responders and 16% for nonresponders. However, suppose that only 20 of the 50 patients who respond do so before the beginning of week 7, with the remaining 30

Table 1. Calculation of Week 7 Death Rates

Method	Die	Survive	Total	Death Rate
Usual Method				
Responders	2	48	50	4%
Nonresponders	8	42	50	16%
Mantel-Byar Method				
Response	2	18	20	10%
No Response	8	72	80	10%

NOTE. For a complete description of this hypothetical example, please see text.

responders entering the response category after week 7. The number of patients in the response category at week 7 is thus 20, of which two die. The number of patients not in the response category at week 7 is 80, of which eight die. Thus, the death rate among patients at risk of dying in week 7 is actually the same for the two groups, namely 10%. The usual method counts survival time before response as time at risk of death for the responder group, thereby underestimating the death rate for responders and overestimating the death rate for nonresponders. As a result of this bias, neither the log rank test nor any other test of statistical significance provides a valid comparison of the risk of death in the two groups. Therefore, such tests should never be used in the usual method to compare survival for responders and nonresponders.

A Modification to the Usual Method

In an attempt to remove some of the bias in the usual method, analyses are sometimes conducted which compare overall survival among nonresponders to survival from time of response among responders. At first glance, this appears to correct the bias caused by the guarantee time since the time to respond is subtracted from the survivorship function of the responders. In fact, such a comparison appears to be conservative in the sense that it is biased in a direction which will favor the nonresponder group.

The modified method, however, may still be biased in favor of responders. Firstly, the death rate for nonresponders will still be overestimated during the early months, since the actual number of patients at risk is larger than the number considered by the method. All of the patients who eventually respond are removed completely from the estimation of the death rate for the nonresponders. The argument is similar to that of the example cited for the usual method.

Secondly, an additional source of bias will affect this analysis if the death rates are changing rapidly over time. For example, assume that the death rates for responders and nonresponders are equal and are as shown in Fig. 1, *top*. Thus, the risk of death for a patient on-study at each time point does not depend on the current response category of that patient. Suppose further that most patients who achieve a response do so at about six months. Then the risk of death for responders will look something like that shown in Fig. 1, *bottom*, if survival time is measured from the onset of response. The death rate function for responders in Fig. 1, *bottom*, is clearly lower than the death rate function for nonresponders in Fig. 1, *top*. The method is likely to show a significant difference beween responders and nonresponders. However, the death rates for responders and nonresponders are, in fact, the same (Fig. 1, *top*), and the difference observed results from comparing different segments of this common death rate function.

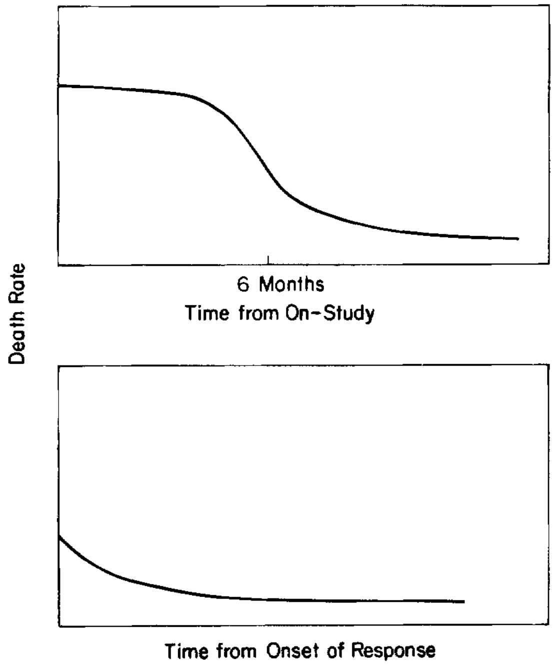

Fig. 1. Death rate functions from a hypothetical example. (*Top*) Death rate function for both responders and nonresponders measured from the start of treatment and (*bottom*) death rate function for responders measured from the start of response.

Mantel-Byar Approach

Neither the usual method nor its modification discussed above are valid for comparing the survival experience of responders to nonresponders. This means that even when response does not confer a survival advantage, the usual method will often falsely conclude that there is a significant difference in survival between the two groups. There are two valid methods for testing the hypothesis that responders have better survival than nonresponders. Neither of these methods provides a test for the effectiveness of the treatment in improving survival.

The first valid method is described in a paper by Mantel and Byar[2] in which the benefit of receiving a heart transplant is evaluated by comparing the risk of death for patients who have not (yet) received a transplanted heart to the risk of death among patients who have been transplanted. The same approach can be applied to the problem of comparing survival between responders and nonresponders. Patients accrue follow-up time to the various "states" which they occupy during treatment. All of the patients begin at on-study time in the "no response" state. Those patients who eventually respond enter the "response" state at the time of response and remain there until death or censoring. At each death time (as measured from on-study), the number of patients in each response state at that time is used to estimate the risk of death for each response category (Table 1). Early deaths in the "no response" group are not unusual as most patients will be in the "no response" category at times soon after study entry. Only later as patients have an opportunity to respond will a continued excess of deaths from the "no response" state provide evidence that death rates do differ depending on response status.

This method assesses whether the risk of death is higher for patients who had not previously attained a response relative to those patients who had. The analysis removes the bias inherent in the usual method, as patients are compared according to their response status at various periods during follow-up. Tumor response status is a covariate which can change over time, thus allowing patients to accrue times at risk of death to the appropriate response category. While such time-dependent covariate analyses are nonstandard, techniques to complete the analysis are available.[2]

The major advantage of the Mantel-Byar method is that it is the most powerful test of the hypothesis of equal death rates for each response category versus the alternative that death rates for the two response states differ by the same proportion over time (proportional hazards). A disadvantage of the method is that it does not produce meaningful survival curves for both responders and nonresponders. Based on the Mantel-Byar model, an estimate of the survival curve for nonresponders can be obtained by considering patients who eventually respond to be censored at the time of response. Note that all study patients are used to estimate the nonresponder survival course. It is not possible to produce a similar survival curve for patients in the response category.

One possible way to show graphically whether survival differs with response category is to contrast the survival curve for nonresponders estimated by the above technique, to the overall survival curve for all patients. If the null hypothesis is true, these two curves should look approximately the same. If the death rate is higher for nonresponders, then the estimated nonresponder survival curve will be below the overall survival curve. Such a presentation may be useful as a display to illustrate whether nonresponders are at greater risk of death than the population taken all together.

The Landmark Method

A second valid method for evaluating survival by tumor response selects some fixed time after the initiation of therapy as a landmark for conducting the analysis. Those patients still on study at the landmark time are separated into two response categories according to whether they have responded before that time. Patients are then followed forward in time to ascertain whether survival from the landmark depends on the patient's response status at the landmark. Patients who go off protocol before the time of landmark evaluation are excluded from the analysis and patients are analyzed according to their response status at the landmark time regardless of any subsequent shifts in tumor response status. Thus, probability estimates and statistical tests are conditional on the response status of patients at the landmark time.

This approach effectively removes the bias present in the usual method and its modification. Unlike the Mantel-Byar method, estimates of survival probabilities as functions of response state at the landmark are available. In addition, a correct statistical test of significance for differences in survival by tumor response can be conducted. The null hypothesis of interest is that survival from landmark does not depend on response status at landmark versus the alternative that subsequent survival does depend on landmark response status. Previously discussed biases have been removed from the analysis and replaced by a conditional approach. A major disadvantage of this method is that the results will depend on the selection of an arbitrary landmark time, and conclusions from the analysis may differ depending on which landmark is chosen. If the choice of landmark is made only after inspection of the data, additional biases are introduced into the analysis. Thus, the selection of landmark should be made before the data analysis, and should be based on some natural time of clinical significance (for example, the end of induction therapy). Another disadvantage of this method is that patients who die before the landmark time do not contribute to the analysis, and patients who respond only after the landmark time are analyzed as nonresponders. Ettinger and Lagakos[3] provide an example of the appropriate application of this method.

The Mantel-Byar and landmark methods credit survival time to the responder category only after a response is observed. However, many investigators believe that biologic activity, as reflected by an observed tumor response, begins to decrease the probability of dying from the start of treatment, not just after a response is observed. In this case the power of the proposed methods to detect a difference in survival will be less than if a test were used which credits some preresponse survival to the responder category. Unfortunately, any such test will be invalid, for the reasons discussed earlier. Therefore the Mantel-Byar and landmark methods may still be the best available valid methods for testing the hypothesis that biologic activity is related to better survival. However, the demonstration of such a relationship cannot be taken as proof that treatment improves survival, even if a valid statistical method is used.

An Example

Data from the Eastern Cooperative Oncology Group (ECOG) Study 3477: Phase II Master Protocol for Evaluation of Agents in Patients with Multiple Myeloma are used to illustrate the proposed methods. Figure 2 shows the survival from on-study time by response for 35 patients evaluable for response treated with cyclophosphamide (600 mg/m^2 on days 1–4 and days 29–32). Responders have a median survival which is over twice as long as the median survival for the nonresponders. It can be clearly seen in Fig. 2 that the difference between the two curves is almost entirely due to the four nonresponders who died early, before most responders had responded. The bias caused by the guarantee time is clearly visible in this example.

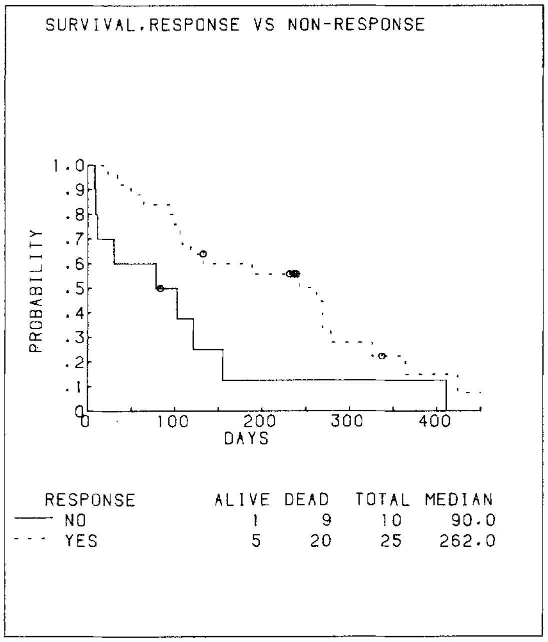

Fig. 2. Evaluation of survival for responders versus nonresponders by the usual method. Data from Eastern Cooperative Oncology Group Study EST 3477: Phase II master protocol for evaluation of agents in patients with multiple myeloma.

The calculation of the chi-square statistic for the log rank test by the usual method is illustrated in Table 2. The idea behind the log rank test is that, under the null hypothesis of equality of survival for responders and nonresponders, the expected number of deaths in a category at a time of death should be proportional to the fraction of all patients at risk who are in that category at that time. If there is no difference in survival between responders and nonresponders, the observed number of deaths among nonresponders (O_{nr}) should be close to the expected number of deaths among nonresponders (E_{nr}). The formulae for the variance of O_{nr} and the chi-square statistic (using Yates' continuity correction) are given at the bottom of Table 2. The *p* value is 0.08, by the usual method. The standard practice in articles which contain an examination of survival by response

Table 2. Evaluation of Survival for Responders Versus Nonresponders by the Usual Method

	Nonresponders			Responders			
Days from On-Study	At Risk (N_{nr})	Deaths (O_{nr})	Expected Deaths (E_{nr})	At Risk (N_r)	Deaths (O_r)	Expected Deaths (E_r)	Variance (V)
8	10	1	0.29	25	0	0.71	0.204
9	9	1	0.26	25	0	0.74	0.195
11	8	1	0.24	25	0	0.76	0.184
23	7	0	0.22	25	1	0.78	0.171
30	7	1	0.23	24	0	0.77	0.175
34	6	0	0.20	24	1	0.80	0.160
49	6	0	0.21	23	1	0.79	0.164
64	6	0	0.21	22	1	0.79	0.168
78	6	1	0.22	21	0	0.78	0.173
94	5	0	0.19	21	1	0.81	0.155
99	5	0	0.20	20	1	0.80	0.160
102	5	1	0.21	19	0	0.79	0.165
106	4	0	0.17	19	1	0.83	0.144
108	4	0	0.18	18	1	0.82	0.149
118	4	0	0.19	17	1	0.81	0.154
121	4	1	0.20	16	0	0.80	0.160
132	2	0	0.11	16	1	0.89	0.099
155	2	1	0.13	14	0	0.88	0.109
188	1	0	0.07	14	1	0.93	0.062
242	1	0	0.09	10	1	0.91	0.083
262	1	0	0.10	9	1	0.90	0.090
269	1	0	0.22	8	2	1.78	0.173
278	1	0	0.14	6	1	0.86	0.122
326	1	0	0.17	5	1	0.83	0.139
364	1	0	0.25	3	1	0.75	0.188
411	1	1	0.33	2	0	0.67	0.222
424	0	0	0.00	2	1	1.00	0.000
746	0	0	0.00	1	1	1.00	0.000
		9	5.04		20	23.96	3.967

NOTE. Data are from the Eastern Cooperative Oncology Group Study EST 3477; Phase II master protocol for evaluation of agents in multiple myeloma. For each time of death compute: $N = N_{nr} + N_r$; $O = O_{nr} + O_r$; $E_{nr} = O\,N_{nr}/N$; and $V = N_{nr}\,O\,N_r\,(N - O)/N^2\,(N - 1)$. To compute sum over death times:

$$\text{chi-square} = \frac{(|\Sigma\, O_{nr} - \Sigma\, E_{nr}| - 0.5)^2}{\Sigma\, V} = \frac{(|9 - 5.04| - 0.5)^2}{3.967} = 3.02,$$

where Σ = summation over all death times and p value = 0.08.

category is to present this p value and survival curves similar to those shown in Fig. 2.

The landmark time analysis in which survival from day 56 is evaluated by response status at day 56 is shown in Fig. 3. Day 56 was the end of the induction phase of this study. The eight patients who died or relapsed prior to day 56 have been excluded from the analysis. The impression given by the curves in Fig. 3 with respect to the effect of response on survival is very different from that given by the curves in Fig. 2. No effect of response status at day 56 is apparent (p = 0.56) now that the bias has been removed.

The Mantel-Byar approach in which patients accrue time in the various states to determine the risk of death from each state is illustrated in Table 3. The calculation of the chi-square statistic in the Mantel-Byar analysis is essentially the same as in the standard log rank test. The only difference is that the number of patients at risk in a given response category can increase as well as decrease with time. When the log rank test is used with a covariate which is not time dependent, the number of patients at risk in a category starts out as a fixed value, and then decreases as patients in that category die or are lost to follow-up (censored). With a time-dependent covariate, the number of patients at risk in a category can also increase due to patients moving from one category to another. In the analysis of survival by response category, this occurs when a patient responds, and hence moves from the nonresponse to the response category. When this happens, the number of patients at risk in the nonresponse group (N_{nr}) decreases by one and the number of patients at risk in the response category (N_r) increases by one.

The difference between the usual and the Mantel-Byar analyses can be seen by comparing Tables 2 and 3. The observed number of deaths in each category is the same in the two tables. However, the expected number of deaths is closer to the observed number of deaths in Table 3. This is due to the fact that for deaths which occur early, the number of patients at risk in the nonresponse category is much higher in the Mantel-Byar analysis than in the usual analysis.

Figure 4 shows survival for all patients, and survival for nonresponders based on the Mantel-Byar analysis. As in Fig. 3, there is little difference between the two curves.

INTERPRETATION OF RESULTS

Implied Causation

A statistically significant difference in survival is frequently interpreted to mean that response causes longer survival. Such a conclusion is generally not justified. Response may act as a mechanism to select a prognostically favorable subgroup of patients, based on the status of the disease or the characteristics of the host, which would have had better than average survival even without the treatment under study. Thus, responders may survive longer than nonresponders, not because of an effect of response on survival, but because response serves to identify patients with pretreatment characteristics which favor longer survival.

It will generally be extremely difficult to differentiate between cases where response prolongs survival and those where it simply acts as a marker for favorable prognosis patients. One approach is to search for covariates which are correlated with both survival and response which might explain some of the results. If a significant relationship between response and survival disappears when these covariates are taken into account (by stratification, for example), then the evidence for a causal relationship between re-

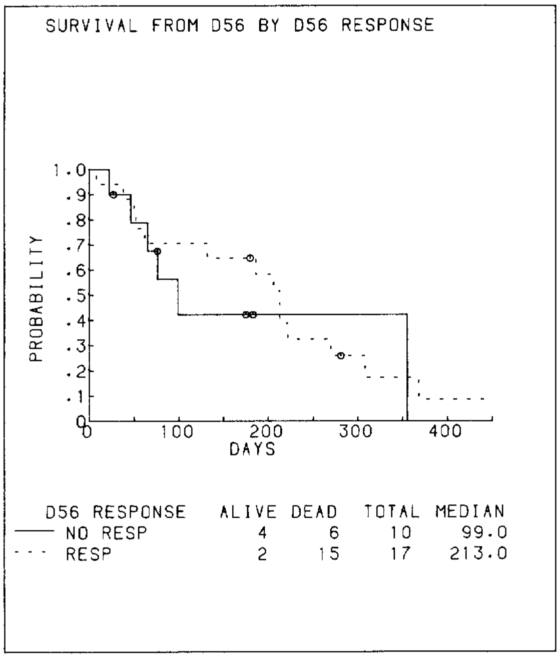

Fig. 3. Evaluation of survival for responders versus nonresponders by the landmark method with landmark at day 56. Data from Eastern Cooperative Oncology Group Study EST 3477: Phase II master protocol for evaluation of agents in multiple myeloma.

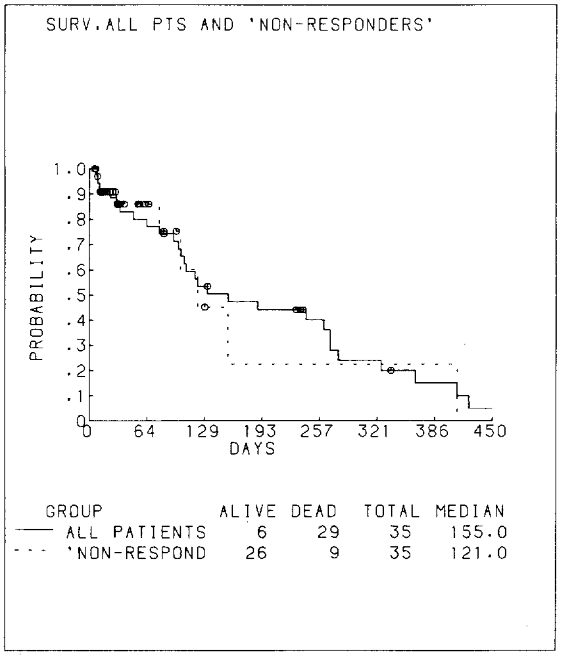

Fig. 4. Evaluation of survival for all patients versus survival for nonresponders by the Mantel-Byar method. Data from Eastern Cooperative Oncology Group Study EST 3477: Phase II master protocol for evaluation of agents in multiple myeloma.

Table 3. Evaluation of Survival for Responders Versus Nonresponders by the Mantel-Byar Method

	Patients in Nonresponse State			Patients in Response State			
Days from On-Study	At Risk (N_{nr})	Deaths (O_{nr})	Expected Deaths (E_{nr})	At Risk (N_r)	Deaths (O_r)	Expected Deaths (E_r)	Variance (V)
8	33	1	0.94	2	0	0.06	0.054
9	31	1	0.91	3	0	0.09	0.080
11	30	1	0.91	3	0	0.09	0.083
23	21	0	0.66	11	1	0.34	0.226
30	19	1	0.61	12	0	0.39	0.237
34	13	0	0.43	17	1	0.57	0.246
49	12	0	0.41	17	1	0.59	0.243
64	9	0	0.32	19	1	0.68	0.218
78	8	1	0.30	19	0	0.70	0.209
94	6	0	0.24	19	1	0.76	0.182
99	5	0	0.21	19	1	0.79	0.165
102	5	1	0.22	18	0	0.78	0.170
106	4	0	0.18	18	1	0.82	0.149
108	4	0	0.19	17	1	0.81	0.154
118	4	0	0.20	16	1	0.80	0.160
121	4	1	0.21	15	0	0.79	0.166
132	2	0	0.11	16	1	0.89	0.099
155	2	1	0.13	14	0	0.88	0.109
188	1	0	0.07	14	1	0.93	0.062
242	1	0	0.09	10	1	0.91	0.083
262	1	0	0.10	9	1	0.90	0.090
269	1	0	0.22	8	2	1.78	0.173
278	1	0	0.14	6	1	0.86	0.122
326	1	0	0.17	5	1	0.83	0.139
364	1	0	0.25	3	1	0.75	0.188
411	1	1	0.33	2	0	0.67	0.222
424	0	0	0.00	2	1	1.00	0.000
746	0	0	0.00	1	1	1.00	0.000
		9	8.56		20	20.44	4.028

NOTE. Data are from the Eastern Cooperative Oncology Group Study EST 3477; Phase II master protocol for evaluation of agents in multiple myeloma. For each time of death compute: $N = N_{nr} + N_r$; $O = O_{nr} + O_r$; $E_{nr} = O\,N_{nr}/N$; and $V = N_{nr}\,O\,N_r\,(N - O)/N^2\,(N - 1)$. Sum over death times:

$$\text{chi-square} = \frac{(|\Sigma\,O_{nr} - \Sigma\,E_{nr}| - 0.5)^2}{\Sigma\,V} = \frac{(|9 - 8.56| - 0.5)^2}{4.028} = 0,$$

where Σ = summation over all death times and p value = 1.00.

sponse and survival is reduced. For example, Besa et al,[4] studied androgen therapy for patients with agnogenic myeloid metaplasia. In that study, patients with normal chromosomes lived significantly longer than patients with chromosome abnormalities ($p < 0.009$), and responders lived longer than nonresponders ($p < 0.005$). In addition, 11 of 13 responders had normal chromosomes, compared to only one of eight nonresponders. Thus, a clearly identifiable favorable subgroup exists which may explain the increased survival among responders.

While inferring that tumor response causes improved survival may seem biologically correct, it is generally not possible to prove causation on the basis of clinical trial data alone. Associations between response and survival may be investigated using the appropriate methods described earlier, but claims that tumor response improves survival should be avoided.

Clinically Useful Measures

Estimates of survival curves for responders and nonresponders can be useful in the clinical management of patients. Response state at a given landmark time can be thought of as a prognostic factor for the future survival of a patient. Poor prognosis patients will supposedly be treated differently than patients with a favorable prognosis. In this situation, the p value of the comparison may be useful to indicate whether the difference in prognosis between responders and nonresponders as defined at the landmark time is merely due to random variations or is statistically significant. The issue of causality is irrelevant when response is used in this way as a prognostic factor.

Validation of Response Criteria

A comparison of survival by response state can be used to validate the prognostic significance of a system for classifying response. An example is provided by Slack et al,[5] in which the response categories ''stable'' and ''progressive disease'' are shown to have different prognoses. Unfortunately, the usual method of analysis is used, although the authors recognize that it is biased.

Response Analysis as a Substitute for a Comparative Trial

The comparison of responders to nonresponders is often interpreted as providing an indication of the therapeutic value of the treatment regimen. Responders are taken as the group which really received the benefit of treatment, as indicated by their response. Nonresponders are taken as a group which did not benefit from the treatment. This latter group is often viewed as a ''control'' group, and their survival, implicitly, taken to be equivalent to the survival for an untreated group. The statistical comparison of responders to nonresponders is thus presented as a surrogate for an analysis of treated versus untreated patients. The ''statistical significance'' of this comparison is often presented as scientific evidence favoring the use of the regimen. This interpretation is clearly wrong. As discussed above, it may be that response to this treatment is a good marker for identifying patients in a favorable subgroup. The best way to determine comparative efficacy of a treatment regimen is to conduct a controlled clinical trial. Additional discussion of this issue and an example are presented in a recent article by Mantel.[6]

The fact that responders live longer than nonresponders is sometimes interpreted to mean that a more aggressive treatment regimen (which produces a higher response rate) will result in longer survival. Again, this may be true but can best be demonstrated by a controlled clinical trial.

RECOMMENDATIONS

In conclusion, we make these recommendations concerning the analysis of survival by response:

(1) The usual method and its modification described earlier are wrong and should never be used. The survival plots produced are biased, the statistical test of significance is invalid, and the conclusions are misleading. This bias is quite large if many patients die early, before the first response evaluation or before most responders achieve a response. Only if most patients who respond do so early and very few patients die in this early period are the results for the incorrect method close to those of the correct methods.

(2) A statistical test of significance for the null hypothesis that the death rate for responders does not differ from that of nonresponders can be performed by considering response state as a time varying covariate using the Mantel-Byar procedure. This procedure provides an effective way to analyze the association between survival and response category, but does not provide estimates of survival probabilities for responders and nonresponders.

(3) Plots of survival by response status, based on the landmark method, can be quite useful descriptively. Given that a patient has or has not responded before a certain landmark time, the prognosis for survival can be determined. This type of conditional information may be useful for the clinical management of patients.

(4) It is wrong to imply that longer survival for responders proves that the treatment being studied is effective, or that a more aggressive treatment is warranted. These comparative questions can best be addressed by a randomized clinical trial.

(5) In every paper submitted for publication which contains a comparison of survival by response, the reason such a comparison is being presented should be clearly stated, and the limitation of the comparison should be discussed. This is necessary to avoid the overinterpretation of such comparisons, and to prevent the drawing of incorrect inferences from the data.

ACKNOWLEDGMENTS

We thank the Eastern Cooperative Oncology Group for the use of the data from study EST 3477. We also acknowledge the helpful comments of Drs. Colin Begg, John Simes, and the reviewers.

REFERENCES

1. Peto R, Pike MC, Armitage NE, et al: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II. Analysis and examples. Br J Cancer 35:1–39, 1977

2. Mantel N, Byar DP: Evaluation of response-time data involving transient states: An illustration using heart-transplant data. JASA 69:81–86, 1974

3. Ettinger DS, Lagakos S: Phase III study of CCNU, cyclophosphamide, Adriamycin, vincristine, and VP-16 in small-cell carcinoma of the lung. Cancer 49:1544 –1554, 1982

4. Besa EC, Norwell PC, Geller NL, et al: Analysis of the androgen response of 23 patients with agnogenic myeloid metaplasia: The value of chromosomal studies in predicting response and survival. Cancer 49:308–313, 1982

5. Slack NH, Mittleman A, Brady MF, et al: The importance of the stable category for chemotherapy treated patients with advanced and relapsing prostate cancer. Cancer 46:2393 –2402, 1980

6. Mantel N: An uncontrolled clinical trial—treatment response or spontaneous improvement? Controlled Clinical Trials 3:369 –370, 1982