

ADJUSTING FOR EARLY TREATMENT TERMINATION IN COMPARATIVE CLINICAL TRIALS

S. W. LAGAKOS, L. L-Y. LIM AND J. M. ROBINS

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

SUMMARY

In clinical trials of long-term therapies, patients often terminate their treatments earlier than planned. When analysing time-to-failure data, one approach to account for early treatment termination censors failure at the time of termination of therapy. In general, however, this does not produce valid inferences about the distribution of time to failure that would have occurred had treatment not been terminated. In contrast, intent-to-treat analyses, which are based on time to failure regardless of whether and when treatment is terminated, always produce valid inferences about the unconditional distribution of time to failure. Early treatment termination does not distort the size (type I error rate) of intent-to-treat tests but can cause a loss in power. Modifications to ordinary logrank tests can be used to recover some of the lost power without affecting test size, and can be most useful when the proportion of at-risk patients still taking their treatment changes substantially during periods when failures are observed. Extensions of the modified test to include strata are straightforward, although important design questions require further research.

INTRODUCTION

Early termination of long-term treatments is a common problem in clinical trials, especially those in which the treatments are self-administered.¹ Reasons for early termination include: the treatment can have side-effects that discourage its use; patients might have the perception that the treatment is not working and thus lose the incentive to take it; deterioration in a patient's condition might cause him or her to stop the treatment; or, the treatment schedule could be so intense that even well-intentioned patients have difficulty complying.

Extensive research has been devoted to issues surrounding early termination of treatment and, more generally, non-compliance with treatment. Much of this addresses ways of preventing non-compliance¹ or ways of measuring and assessing the extent of non-compliance that occurs in a trial.^{2,3} Relatively little attention has been given to assessing the effects of early termination on treatment comparisons or on ways of adjusting statistical tests to account for early treatment termination. Detre and Peduzzi⁴ analyse the baseline distributions of covariates in compliant and non-compliant patients to assess the possibility that non-compliance may have biased the treatment comparison. Newcombe⁵ proposes a method of adjusting estimates of treatment effects that are based on knowledge about the extent of non-compliance. Robins^{6,7} discusses limitations of existing methods and proposes methods for correcting for non-compliance based on causal inference models.

The first objective of this paper is to describe the effects of early treatment termination on statistical tests for comparing treatment groups in a randomized clinical trial in which the outcome variable is the time until some event of interest, such as survival time. We consider two

approaches: (1) analyses which censor survival at the time of treatment termination, and (2) intent-to-treat analyses which utilize the complete observation of survival time, regardless of whether and when treatment is terminated. For each we assess the effects of non-compliance on statistical comparisons of the treatment groups being examined. A second objective is to examine an approach that attempts to minimize the negative effects of non-compliance by modifying standard tests. Several related problems, such as extensions of the methods and design implications, are also discussed.

NOTATION AND SETTING

Consider a randomized clinical trial in which two groups are compared with respect to some failure time, and in which the treatments are long term – for example, a self-administered medication that is intended to be taken until the time of failure. Suppose, however, that some patients stop taking their treatments before they fail. To describe this process for someone in one of the two treatment groups, let U denote the time until the earlier of failure or termination of treatment, and let ε be an indicator of whether U corresponds to time until failure ($\varepsilon = 1$) or treatment termination ($\varepsilon = 0$). Also, let T denote time until failure, regardless of whether or not treatment is terminated. Then the probabilistic aspects of (U, ε, T) can be described by functions

$$h_d(u) = \lim_{\Delta \rightarrow 0} Pr(U < u + \Delta, \varepsilon = 0 | U \geq u) / \Delta$$

$$h_{f|c}(u) = \lim_{\Delta \rightarrow 0} Pr(U < u + \Delta, \varepsilon = 1 | U \geq u) / \Delta$$

$$h_{f|d}(t|u) = \lim_{\Delta \rightarrow 0} Pr(T < t + \Delta | T \geq t, U = u, \varepsilon = 0) / \Delta.$$

The functions $h_d(u)$ and $h_{f|c}(u)$ are the cause-specific hazard functions⁸ for terminating treatment at time u and for failing while continuing on treatment at time u , respectively. For individuals who terminate treatment at time u , $h_{f|d}(t|u)$ is the hazard function for failing at time t . We denote the unconditional hazard function for failing at time t by $h_f(t)$, that is

$$h_f(t) = \lim_{\Delta \rightarrow 0} Pr(T < t + \Delta | T \geq t) / \Delta.$$

An expression for h_f in terms of h_d , $h_{f|c}$ and $h_{f|d}$ is given in Lagakos and Ryan.⁹

TWO METHODS OF ANALYSIS

We consider two approaches for analysing failure time data in the setting described above. Both approaches define a 'survival time' and 'censoring indicator' for each patient, but they differ in how they handle patients that terminate treatment early. For simplicity of presentation, we ignore end-of-study censoring. However, the adaptation of either method to account for this is straightforward.

Approach 1: censoring failure at the time of treatment termination

Consider first an analysis in which 'survival time' is taken to be the earlier of time to failure and time to treatment termination (U), and is regarded as being censored if it represents time to treatment termination. The rationale behind this approach is that once patients terminate treatment, their survival times no longer fully reflect the benefits that could be provided by that treatment. Thus, it would seem to make sense to censor their survivals at the time of treatment termination.⁴

Without further assumptions, application of usual survival analysis methods, such as the Kaplan-Meier estimator, to these data produces inferences about the cause-specific hazard function $h_{f|c}(u)$; that is, the hazard of failing while a patient in the trial is still continuing on treatment. Yet $h_{f|c}$ is not, in general, the hazard function, say $h^*(t)$, that would have been experienced had patients not terminated treatment. These would be equal only if those patients who terminate treatment at time u do not differ from those who continue on treatment at time u with respect to any measured or unmeasured covariates.⁶ In most situations, this condition will not hold because patients who terminate treatment would be 'different' from those who continue treatment. Furthermore, without any additional assumptions, $h^*(t)$ is non-identifiable; that is, it cannot be expressed in terms of the identifiable functions $h_{f|c}$, h_d and $h_{f|d}$. The situation is analogous to one in competing risks in which one would like to know the effects on survival of eliminating a particular cause of death.⁸

Given that the conditions needed for $h_{f|c}$ to equal h^* are not met, the appropriateness of $h_{f|c}$ as a basis for assessing and comparing the treatment groups comes into question. For example, if patients tend to terminate treatment as their condition begins to deteriorate, this approach would censor their survival times before their imminent failure. In general, it is not appropriate to assess the relative efficacy of the treatment groups by comparing the functions $h_{f|c}$.

Approach 2: intent-to-treat analysis

The logical and statistical problems associated with approach 1 seem to be widely appreciated in statistical practice, yet relatively little discussion appears in the literature. Currently, standard practice is to employ an intent-to-treat analysis in which the observed time until failure (T) is analysed regardless of whether treatment is terminated. In such an analysis, the parameter being estimated or tested is $h_f(t)$, the unconditional hazard function corresponding to failure at time t .

One rationale for this approach is that patients in the trial who terminate their treatments do so for the same reasons that would apply were the treatment to become part of routine medical practice. For example, if some patients in the trial terminate treatment due to toxicity, this would also be expected to occur in routine medical practice. Thus, the hazard function h_f estimated from the results of the trial could be expected to reflect the benefit from use of the treatment in practice. Another reason that this approach is used relates to the inability of the first approach to produce valid inferences about the function h^* . However, even if h^* were estimable, h_f might still be a more appropriate basis for comparison because some reasons for treatment termination may not be avoidable, and thus h^* may not be achievable in practice.

Despite the advantages of intent-to-treat analyses over those based on approach 1, they are not without their own limitations. In particular, if beliefs and circumstances surrounding the use of a treatment change when it is introduced into standard medical practice, then the function $h_d(u)$ and hence the function $h_{f|d}(t|u)$ could also change. If so, the function $h_f(t)$ being estimated in the clinical trial might not reflect what would be realized in general medical practice. This difficulty is not specific to intent-to-treat analyses, but applies to all inferences drawn from clinical trials.

Because of the advantages of intent-to-treat analyses, the remainder of this paper will be devoted to the effects of treatment termination on the efficacy of tests based upon such analyses.

ATTENUATED TREATMENT EFFECTS FROM EARLY TERMINATION

Suppose we wish to compare two treatments, A and B. For simplicity, let treatment A denote some biologically active drug and treatment B denote placebo, and suppose that early termination only applies to the active drug. Let h_d^j , $h_{f|c}^j$, $h_{f|d}^j$ and h_f^j denote the analogues of the functions h_d , $h_{f|c}$, $h_{f|d}$ and h_f for treatment group j , for $j = A, B$.

To see the effects of treatment termination under a specific set of circumstances, suppose that patients in the trial terminate treatment A at random and that the effects of termination are to alter their failure hazard function to that of the placebo group. That is, suppose that the failure hazard function at time t for a patient in group A terminating treatment at time u , that is $h_{f|c}^A(t|u)$, equals $h_{f|c}^B(t)$ ($= h_f^B(t)$). Then

$$h_f^A(t) = \gamma(t)h_{f|c}^A(t) + [1 - \gamma(t)]h_f^B(t), \quad (1)$$

where

$$\gamma(t) = Pr\{U \geq t | T \geq t\} \quad (2)$$

is the conditional probability of being on treatment at time t , given that failure has not yet occurred. Thus, for this special situation, the hazard function for failure at time t in patients assigned to treatment A is a mixture of their cause-specific hazard for failing while on treatment and the hazard for failing in the placebo group.

Now consider the following treatment hazard ratios:

$$\begin{aligned} \rho_{f|c}(t) &= h_{f|c}^A(t)/h_{f|c}^B(t) \\ \rho_f(t) &= h_f^A(t)/h_f^B(t). \end{aligned} \quad (3)$$

Using the assumptions in this example, $\rho_{f|c}$ describes the potential benefit of treatment A if treatment termination could somehow be avoided, whereas ρ_f describes the actual benefit of treatment if termination is not prevented. It follows from (1) and (2) that

$$\rho_f(t) = \gamma(t)\rho_{f|c}(t) + 1 - \gamma(t). \quad (4)$$

Thus, if $\rho_{f|c}(t) \leq 1$, then $\rho_{f|c}(t) \leq \rho_f(t) \leq 1$; that is, the effects of treatment termination are to attenuate the treatment difference from $\rho_{f|c}$ to ρ_f .

WEIGHTED LOGRANK TESTS

Suppose that we wished to employ an intent-to-treat analysis to preserve test validity, but that we also suspected that equation (4) held, at least approximately. Then, as has been proposed by Robins,⁶ one can attempt to improve ordinary survival analyses by designing tests that are optimal for $\rho_f(t)$ given some assumptions about $\rho_{f|c}(t)$. To illustrate, suppose that $\rho_{f|c}(t) = \theta$ for some unknown proportionality constant θ . Then equation (4) reduces to

$$\begin{aligned} \rho_f(t) &= \theta\gamma(t) + 1 - \gamma(t) \\ &= (\theta - 1)\gamma(t) + 1. \end{aligned} \quad (5)$$

Thus, rather than utilize an ordinary logrank test in an intent-to-treat analysis, one could use a weighted logrank test with weights selected to reflect the induced changes in $\rho_f(t)$ with time. The ordinary logrank test statistic can be expressed as^{10,11}

$$\sum(O_j - E_j)/\{\sum V_j\}^{1/2},$$

where the sum is over the observed failure time t_j ; O_j and E_j are the observed and expected numbers of failures at t_j ; and V_j is the variance of $(O_j - E_j)$. A weighted logrank test is given by¹²

$$\sum w_j(O_j - E_j)/\{\sum w_j^2 V_j\}^{1/2},$$

where the w_j are some weights. Schoenfeld¹² has shown that the optimal logrank test would use weights proportional to the logarithm of $\rho_f(t_j)$; that is, $w_j \propto \log\{\rho_f(t_j)\}$. Under our assumptions,

it follows from (5) that

$$w_j = \log[\theta\gamma(t_j) + 1 - \gamma(t_j)]. \quad (6)$$

Thus, one can postulate a model for how the risk of failure changes with the termination of treatment, and from this express the induced hazard ratio $\rho_f(\cdot)$ in terms of the pure hazard ratio. The key point is that regardless of what model is assumed, any resulting test of the hypothesis $H_0: \rho_f^A(\cdot) = \rho_f^B(\cdot)$ is valid when an intent-to-treat analysis is used. Use of an incorrect assumption about the effects of termination on the risk of failure will only affect the efficiency of the resulting test.

To apply this approach, it is necessary to know or estimate the parameters θ and $\gamma(t)$. We return to this in a later section, after first considering the potential gains in efficiency from using a weighted versus ordinary logrank test. We also note that use of weighted logrank tests is also appropriate in situations where the postulated model is not proportional hazards; see, for example, Self *et al.*¹³

EFFICIENCY GAINS FROM WEIGHTED LOGRANK TESTS

In this section we investigate the potential for improving efficiency through the use of a weighted versus unweighted logrank test. We wish to assess whether, in practice, important gains in efficiency can be expected and in what circumstances. Using the results in Schoenfeld,¹² we computed the asymptotic relative efficiency (ARE) of the ordinary logrank test to the optimal weighted logrank test for a variety of choices for θ , $\gamma(t)$, $h_{f|c}$, and the degree of end-of-study censoring. The latter is introduced by assuming that patients in the trial accrue uniformly in the chronologic period $[0, t_1]$, and that the data are analysed at chronologic time t_2 ($t_1 \leq t_2$). We considered functions γ which decline linearly from $\gamma(0) = 1$ to the value $\gamma(\tau)$ at time τ , and are constant thereafter. Finally, we assumed that $h_{f|c}$ corresponds to exponential survival with mean μ . We evaluated the ARE for the 1620 situations generated by the following values for θ , $\gamma(\tau)$, τ , μ and (t_1, t_2) , which would describe the conditions of many clinical trials:

$$\theta = 1.3, 2, 3, 5$$

$$\gamma(\tau) = 0.2, 0.4, 0.6$$

$$\tau = 1, 2, 3$$

$$\mu = 0.5, 1, 2, 4, 8$$

$$(t_1, t_2) = (0.5, 1), (0.5, 2), (0.5, 3), (1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3).$$

We note that not all of these 1620 situations are distinct. Specifically, the results for $(\tau, \gamma(\tau)) = (3, 0.4)$ are the same as those for $(2, 0.6)$ when $t_2 = 1$ or 2, and the results for $(\tau, \gamma(\tau)) = (2, 0.2)$ are the same as those for $(1, 0.6)$ when $t_2 = 1$. Details of these calculations are available upon request. Here we only summarize the results. In all the calculations, the ARE varied very little with θ . In 91 per cent of the cases, the AREs are at least 90 per cent. Thus, despite the induced changes in the treatment hazard ratio, the ordinary logrank test usually lost little efficiency relative to the optimal weighted logrank test. This is not surprising since the ordinary logrank test is known to maintain high efficiency as long as the treatment hazard does not vary greatly during periods when failures are observed.¹⁴ We examined in greater detail the 139 situations in which the ARE was below 90 per cent. The poorer performance of the ordinary logrank test tended to occur with large values of t_2 and with smaller values of τ , μ and γ ; that is, in situations where γ changes substantially during periods where many failures were observed. For

example, in a trial with one year of accrual and two additional years of follow-up, a mean survival of 4 years, a true hazard ratio of 2, and a γ that declines to 0.20 at one year, the ARE of the logrank test relative to the weighted logrank test is 77 per cent.

Further AREs were calculated to determine whether an approximation to $\gamma(t)$ would lose much efficiency. Because weights are a function of θ and γ , we used for our selection of non-optimal weights the 36 combinations of θ , γ and τ given above. For each of the situations in which the ARE was below 90 per cent, we computed the AREs of the other 35 weighted logrank test with weights based on these 35 sets of parameters to the optimal weighted logrank test. The results indicated that with this choice of approximate weights the weighted logrank test maintained very high efficiency against the optimal logrank test. These results suggest that use of approximate weights will lead to very high AREs relative to the optimal weighted test.

ESTIMATION OF THE FUNCTION $\gamma(t)$

In order to use the weights in (6), it is necessary to know or estimate both θ and $\gamma(t)$. Based on the calculations in the preceding section, any plausible value of θ would suffice, and the efficiency of the resulting weighted logrank test is high as long as $\gamma(t)$ is estimated reasonably well. Estimation of $\gamma(t)$ can be made in several ways. For example, in clinical trials of patients with AIDS, a pharmacologic assay exists that will detect levels of the drug AZT from samples of patient sera. Thus, if all or a subset of the patients assigned to AZT are tested for the presence of AZT at varying times after the start of treatment, then $\gamma(t)$ can be estimated.

Robins and Tsiatis¹⁵ show that the usual asymptotic distributional properties of a weighted logrank test still hold when the function $\gamma(t)$ is estimated from the clinical trial in the way described above. However, guidelines for when and how often to sample patients to estimate γ are not straightforward. Because these assays can be expensive, an important research question is to determine when these samples should be taken and how large they need to be. It would seem that adaptive sampling techniques could be helpful in this setting.

DISCUSSION

We have considered the consequences of early termination of long-term treatments in comparative clinical trials. In particular, we focused on the effects of termination on comparisons of treatment groups based on two types of analyses – those in which failure times are censored at the time a patient terminates protocol treatment, and those using an intent-to-treat analysis. The first approach does not, in general, provide valid inferences about the hazard function for failing had patients not terminated treatment early, but rather about $h_{r|c}$, the cause-specific hazard for failing while on treatment. Moreover, because the patients that terminate treatment early may be different from those who remain on treatment, the function $h_{r|c}$ is usually an inappropriate basis for comparing treatments.

Intent-to-treat analyses avoid these problems by focusing attention on the unconditional hazard function of failure. They lead to valid tests of treatment equality under the null hypothesis, whether or not early termination occurs. Consequently, this approach is preferable to that based on censoring failure times at the time of termination of therapy. The presence of early termination can reduce the statistical power to detect real differences, and we have discussed one approach to help to minimize this loss.

In one special case, the effects of early termination of treatment were to induce non-proportional hazards between the two treatments. The inefficiency of the ordinary logrank test can be improved in these situations by use of a weighted logrank test, with weights dependent

upon the amount and pattern on early termination. In many of the specific examples we investigated, however, the efficiency gains were small. In using a weighted test, it is important to specify in advance of the analysis the specific approach that will be taken. Weight functions which are selected *post hoc* should be viewed with scepticism if their choice was influenced by the significance levels they produce.

The results described in this paper can be generalized in several ways. First, we considered only settings in which an active treatment is compared with a placebo. Similar modifications to the ordinary logrank test can be applied when comparing two active treatments. One need only postulate a model for how the risk of failure in each group changes when treatment is terminated and derive the relationship between $\rho_{f|c}$ and ρ_f . The accuracy of the postulated model does not affect the validity (type I error rate) of the test, but only its power when H_0 is not true.

A second extension, applicable to multi-centre clinical trials, is to allow the function γ to vary with institution. This would be appropriate, for example, when one expects the rates of early treatment termination to vary geographically, as they might in AIDS clinical trials. Zelen¹⁶ has proposed similar modifications for binary outcome data. To accommodate institutional differences, or any other stratification factor, one could use a stratified weighted logrank test, with the weights for a particular institution based on its γ function. That is, one could compute a weighted observed-minus-expected statistic for each institution, sum these, and then standardize the sum by the square root of the sum of the corresponding variance estimates.

Finally, we note that the methods discussed in this paper have focused on questions of testing. It also may be of interest to estimate the pure hazard ratio from the induced hazard ratio in order to get a sense of the potential benefit of therapy if reasons for treatment termination could be corrected. We caution, however, that unlike the results for testing, the validity of any methods for adjusting the observed hazard ratio h_t would depend on correctly specifying how the risk of failure changes upon termination of treatment. Thus, any resulting analyses would need to be regarded as speculative.

ACKNOWLEDGEMENTS

This paper is based on a talk presented at the 1989 ENAR statistical meeting in Lexington, Kentucky. We are grateful to Susan Ellenberg for inviting us to present this paper, to Lynne Billard for her discussion of it, and to the reviewers and editor for their comments. This work was supported by awards AI24643 and AI95030 from the National Institute of Allergy and Infectious Diseases.

REFERENCES

1. Goldman, A. I., Holcomb, R., Perry, H. M., Schnaper, H. W., Fitz, A. E. and Frohlich, E. D. 'Can dropout and other noncompliance be minimized in a clinical trial?', *Controlled Clinical Trials*, 3, 75-89 (1982).
2. Goldsmith, C. H. 'The effect of compliance distributions on therapeutic trials', in Haynes, R. B., Taylor, D. W. and Sackett, D. L. (eds), *Compliance in Health Care*, Johns Hopkins University Press, Baltimore and London, 1979.
3. Gordis, L. 'Conceptual and methodologic problems in measuring patient compliance', in Haynes, R. B., Taylor, D. W. and Sackett, D. L. (eds), *Compliance in Health Care*, Johns Hopkins University Press, Baltimore and London, 1979.
4. Detre, K. and Peduzzi, P. 'The problem of attributing deaths of nonadherers: the VA coronary bypass experience', *Controlled Clinical Trials*, 3, 335-364 (1982).
5. Newcombe, R. G. 'Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur', *Statistics in Medicine*, 7, 1179-1186 (1988).
6. Robins, J. M. 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies', in Sechrest, L., Freeman, H. and Mulley, A. (eds), *Health Service Research Methodology: A Focus on AIDS*, U.S. Public Health Service, 1989, pp. 113-159.

7. Robins, J. M. 'Correcting for noncompliance in randomized trials using rank-preserving structured failure time models', Technical Report, Occupational Health Program, Harvard School of Public Health, Boston, MA, 1989.
8. Prentice, R. L., Kalbfleisch, J. D., Peterson, A., Flourney, N., Farewell, V. and Breslow, N. 'The analysis of failure time in the presence of competing risks', *Biometrics*, **34**, 541-554 (1978).
9. Lagakos, S. W. and Ryan, L. M. 'Statistical analysis of disease onset and lifetime data from tumorigenicity experiments', *Environmental Health Perspectives*, **63**, 211-216 (1985).
10. Mantel, N. 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports*, **50**, 163-170 (1966).
11. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised controlled trials requiring prolonged observation of each patient. I. Introduction and design', *British Journal of Cancer*, **34**, 585-612 (1976).
12. Schoenfeld, D. 'The asymptotic properties of nonparametric tests for comparing survival distributions', *Biometrika*, **68**, 316-319 (1981).
13. Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S. L., Clifford, C., Goldman, S., Urban, N., Sheppard, L. and Greenwald, P. 'Statistical design of the Women's Health Trial', *Controlled Clinical Trials*, **9**, 119-136 (1988).
14. Lagakos, S. W. and Schoenfeld, D. A. 'Properties of proportional-hazards score tests under misspecified regression models', *Biometrics*, **40**, 1037-1048 (1984).
15. Robins, J. M. and Tsiatis, A. A. 'A large-sample study of G-tests and estimators', Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston, MA, 1989.
16. Zelen, M. 'Response', *Journal of Chronic Diseases*, **37**, 954-955 (1984).