STATISTICS IN MEDICINE, VOL. 16, 833–839 (1997)

AN OVERVIEW OF STATISTICAL METHODS FOR MULTIPLE FAILURE TIME DATA IN CLINICAL TRIALS

L. J. WEI* AND DAVID V. GLIDDEN

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

SUMMARY

In a long term clinical trial to evaluate a new treatment, quite often each study subject may experience a number of 'failures' that correspond to repeated occurrences of the same type of event or events of entirely different natures during his/her follow-up period. To obtain efficient inference procedures for the therapeutic effect over time, it is desirable to utilize those multiple event times in the analysis. In this article, we review some useful procedures for analysing different kinds of multivariate failure time data. Specifically, we discuss the two-sample problems and the general regression problems with various survival models. We also give some recommendations of appropriate procedures for each type of multiple event data structure for practical usage. © 1997 by John Wiley & Sons, Ltd. Stat. Med., Vol. 16, 833–839 (1997).

(No. of Figures: 0 No. of Tables: 0 No. of Refs: 37).

1. INTRODUCTION

In a long term clinical comparative study, life is relatively simple for a statistician if the study has a unique well-defined outcome variable to evaluate the treatment difference. When there are a number of equally important endpoints involved in the trial, issues about the design of the study, the analysis of the data, and the interpretation of the results become rather challenging. These multiple comparisons problems have been addressed extensively in the statistical literature[1] and are taken seriously by the drug industry and regulatory authorities.[2]

In this paper, we will not discuss the philosophical issues of the multiplicity problem. Our goal is to review and discuss the existing statistical methods which specifically deal with the multiple failure time data. This type of data is commonly encountered in clinical trials for treating chronic diseases. For such clinical studies, the subject may experience a number of failures which correspond to repeated occurrences of the same type of event or to the occurrence of events of entirely different natures. In the discussions, we classify these multiple failure time data into three categories. In Section 3, we summarize methods for the analysis of the recurrence data or multi-state failure time data. In Section 4, we deal with the case with unordered and distinct events. In Section 5, inference procedures are discussed for highly stratified (or clustered) failure times. For the benefit of our readers, we briefly summarize methods for handling the usual multiple endpoints in Section 2. These methods are essential for the discussions in the later sections. An excellent review on the Cox regression analysis[3] of multivariate failure time data is given by Lin.[4]

* Correspondence to: L. J. Wei.

CCC 0277–6715/97/080833–07$17.50

2. ANALYSIS WITH MULTIPLE ENDPOINTS

Suppose that in a two-arm comparative study, data are collected on K outcome variables from each patient and the treatment difference is evaluated based on those multiple endpoints. For example, in a trial to evaluate if a six month course of oral antioxidants would slow the progression of amyotrophic lateral sclerosis (ALS), the assessment of the ALS progression can be based on bulbar function, respiratory function, muscle strength, and upper and lower extremity function. For each subject, one may construct a single numerical score (for example, Appel rating scale: 0–160) based on the above endpoints and use these global scores to perform univariate analyses for the treatment difference. Although such analyses are conceptually simply, it is more informative to use a multivariate analysis approach to identify the effect of antioxidants on particular endpoints. Here, we briefly describe a general approach that a frequentist might take to handle this multiple comparisons problem.

Let θ_k be the treatment difference with respect to the kth endpoint and let the null hypothesis H_k be $\theta_k = 0$, that is, there is no difference between the two treatment groups with respect to the kth response. Also, let T_k be the corresponding standardized two-sample test statistic (for example, t-test, Wilcoxon test, the difference of two-sample proportions *et al.*), $k = 1, \ldots, K$. Based on those T_k's, one would like to know which θ_k's are not zero. Now, let H be the hypothesis which is the intersection of all the *true* H_k's. Then, the *experimentwise* type I error probability of a test procedure for testing $\{H_k, k = 1, \ldots, K\}$ is the chance of rejecting H. To maintain this prespecified experimentwise type I error rate α one may use the Bonferroni correction.[5] However, this approach is conservative when these statistics T_k's are highly correlated or very few θ_k's are 0. Generally, for a large sample, the joint null distribution of T_k's can be approximated by a multivariate normal distribution with mean 0 and covariance matrix Γ. To relax the conservatism of Bonferroni correction due to the dependence among the test statistics T_k's, one may obtain a cut-off point c using this normal approximation, that is,

$$P\left(\max_{k=1,\ldots,K} |Z_k| < c\right) = 1 - \alpha$$

where $Z = (Z_1, \ldots, Z_K)'$ is the multivariate normal with mean 0 and covariance matrix Γ. Then, if $|T_k| > c$, one may claim that $\theta_k \neq 0$. The critical value c can be obtained by simulating Z. A similar procedure can be derived for testing $\{H_k, k = 1, \ldots, K\}$ against a one-sided alternative hypothesis.

One may make the above testing procedure even less conservative with a sequential method.[6,7] Suppose that $\{|t_1| < \cdots < |t_K|\}$ are observed ordered $\{|T_k|\}$. If $|t_K| > c$, we claim that $\theta_K \neq 0$. We then obtain another cut-off point c_{K-1} such that

$$P\left(\max_{k=1,\ldots,(K-1)} |Z_k| < c_{K-1}\right) = 1 - \alpha.$$

If $|t_{K-1}| > c_{K-1}$, we conclude that $\theta_{K-1} \neq 0$. Testing continues until $|t_l| < c_l$, and one would conclude that $H_{k'}$ is non-significant for $k' \leqslant l$. This multi-stage method preserves the experimentwise type I error probability α and is more powerful than the non-sequential one, especially when very few H_k's are true.

If the treatment differences with respect to all the endpoints are expected to be in the same direction, one may combine the T_k's linearly to create a global assessment of the relative merit of the treatments.[8,9] That is, we consider a statistic $T = \sum_{k=1}^{K} w_k T_k$, where w_k can be chosen to have maximum power to detect a prespecified alternative hypothesis. A common choice of $w = (w_1, \ldots, w_K)'$ is $\Gamma^{-1} e$, where $e' = (1, \ldots, 1)$, a $K \times 1$ vector. Note that except for the sequential

method, all the above procedures can be easily adapted to obtain joint confidence regions or intervals for θ_k's by inverting the corresponding test statistics.

Now, suppose that one needs to make adjustments from some potentially important covariates to make inferences about the treatment difference. When the measurements of all the endpoints are in the same scale, one may use the mixed effects model to estimate the regression coefficients.[10] To relax the parametric assumptions in the mixed effects model, one may take the generalized estimating equations approach[11] to tackle the problem with multiple endpoints.

3. ANALYSIS OF RECURRENCE DATA OR MULTI-STATE FAILURE TIMES

A common feature for the recurrence times data and multi-state failure time observations is that the events are naturally ordered and occur in a certain sequence over time. First, we use a bladder tumour study conducted by the Veterans Administration Cooperative Urological Research Group (VACURG)[12] to illustrate the existing procedures to handle recurrence times data. In this study all the patients had superficial bladder tumours when they entered the trial. These tumours were removed transurethral and patients were randomly assigned to one of the three treatments: placebo; thiotepa, or pyridoxine. Most patients had multiple recurrences of tumours during the study. The new tumours were removed at each visit. The recurrence times of tumours for patients in placebo and thiotepa groups are reported in Wei *et al.*[13] Each recurrence time was measured from the beginning of the patient's treatment. We assume that a patient's follow-up time is independent of his/her event times. Furthermore, for each patient, some potentially important covariates were also recorded, for example, the number of tumours and the size of the largest tumour at patient's entry. To evaluate the thiotepa effect, one may conduct the usual survival analysis based on the first recurrence times. This univariate procedure, however, may not be efficient and cannot be used to examine if there is a temporal treatment effect from thiotepa.

To perform a multivariate analysis, first let us consider two-sample non-parametric tests for testing the equality of the two joint distributions for the first K tumour recurrence times without making covariate adjustments. To mimic the non-censored case discussed in Section 2, for each type of tumour recurrence (the first, the second etc.), we construct a two-sample distribution-free test (for example, the logrank, Gehan test, or more generally, any member of the Gill's class of statistics[14]). This gives us K correlated test statistics, whose null distribution is approximately normal with mean 0. The limiting covariance matrix can be easily estimated.[15] One can then follow the multiple comparisons procedures discussed in Section 2 to make inferences about the differences of two groups. For example, one may be able to claim that thiotepa works well with respect to the first two tumour recurrences, but not so for the later recurrences. To have a global assessment of the treatment effect, a one-degree of freedom test can also be obtained by combining these K statistics linearly.[7,8]

Next, let us consider the regression problem with recurrence times data, where patient's treatment indicator, number of tumours and size of the largest tumour at randomization are the covariates. For this type of data, there are several inference procedures available for the regression coefficients in the literature. First, one may use the Andersen-Gill (AG) model to handle recurrence times data.[16] This model is a generalization of the Cox proportional hazards model and relates the intensity function of tumour recurrences to the covariates multiplicatively. Under this model, the risk of a recurrent event for a patient follows the usual proportional hazards assumption and is unaffected by earlier events that occurred to the patient. The AG model essentially assumes that the tumour recurrences follow a non-homogeneous Poisson process. Such strong Markovian assumptions can be relaxed by introducing time-dependent covariates in the model, such as the number of prior recurrences, which may capture the

dependence structure among the recurrence times. The AG model can be easily analysed by the partial likelihood principle.

Prentice, William and Peterson (PWP)[17] proposed an alternative approach for analysing recurrence data. The PWP model specifies that the hazard function at time t for the kth recurrence of a patient, conditional on the entire failure, censoring and covariate history prior to time t has the usual proportional hazards form, where the nuisance hazard functions may vary with k, $k = 1, \ldots, K$. This produces a proportional hazards model with time dependent strata, where the dependence between event times is handled by stratifying by the prior number of failures. The PWP model can also be analysed by the partial likelihood principle. It has been shown that the AG and PWP models are sensitive to misspecification of the dependence structure among the recurrence times.[13]

The third approach one may take for the present situation is to use the proposal by Wei, Lin and Weissfeld (WLW).[13] Their idea is quite simple, that is, for the kth recurrence times, one uses the usual proportional hazards model (the nuisance hazard function and the regression parameters may vary with k) to obtain the maximum partial likelihood estimate, say, $\hat{\beta}_k$, for the treatment difference. Then, the joint distribution of $\{\hat{\beta}_k, k = 1, \ldots, K\}$ is approximately normal[11] with a covariance matrix which can be consistently estimated. Inferences about the treatment differences can then be made in the spirit of methods mentioned in Section 2. A global evaluation of the treatment difference may be obtained by combining these $\hat{\beta}_k$'s linearly. Recently, Prentice and Cai[18] have tried to obtain more efficient estimation procedures than the WLW method. However, they find that the WLW performs quite well for almost all the practical situations.

The WLW method can be easily adapted to the competing risk setting. If the onset of an event, for example, death, precludes the development of the kth tumour recurrence, the WLW can model the cause-specific hazard for this recurrence, but model the ordinary (net) hazard for death. By doing this, one may obtain a global assessment of the treatment difference with respect to tumour recurrence and patient's mortality. An alternative way to handle the above situation is to redefine the kth event to be either the kth tumour recurrence or death, $k = 1, \ldots, K$.[19]

For the AG, PWP and WLW methods, computer software is available, for example, MUL-COX2,[20] the SAS PHREG procedure, and S macros (developed by Terry Therneau of the Mayo Clinic, Rochester, Minnesota, U.S.A.).

Although the Cox model has been used extensively to examine the covariate effects on the hazard function for the failure time variable, it may not fit the data well. A useful alternative to the proportional hazards model is the so-called accelerated failure time (AFT) model, which simply regresses the logarithm of the survival time over the covariates.[21–25] Using the idea of generalizing the usual univariate linear model to the multivariate case, this log-linear model can be easily extended to multiple event times data.[26] However, it is not clear how this multivariate AFT model can handle the case when there are competing risks.

The fifth method for analysing recurrence time data is to use the so-called frailty model, for example, an Andersen–Gill model with a random effect. The frailty may be thought of as a random covariate which induces dependence among the multiple event times. Conditional on this random effect, for each patient, his/her intensity function for tumour recurrences follows an AG model. The most popular frailty model in survival analysis is the one proposed by Clayton and Cuzick.[27] This model assumes that the frailty is from a gamma distribution with mean one and an unknown variance. Parameter estimation for such a model, however, is difficult since standard partial likelihood methods do not eliminate the nuisance hazard function. Recently, Nielsen *et al.*[28] have proposed an estimation procedure for the regression parameters, the variance of the frailty, and the underlying intensity function. However, their method is computationally demanding and the large sample properties of their estimators are available for very

special cases.[29,30] Moreover, an unpleasant feature of the 'gamma frailty' model is that while hazards are proportional conditional on the frailty, the marginal hazards are not proportional. Hougaard[31] suggests the use of a 'positive stable' distribution for the frailty with the Cox model. This particular distribution guarantees that both conditional and marginal hazards satisfy the proportional hazards assumption. In general the frailty models have an intuitive appeal and provide insight into the relationship between failures. On the other hand, the robustness of such models to misspecification of the frailty distribution is unclear and inference procedures with frailty distributions other than the one by Clayton and Cuzick[27] have not been developed.

Note that the PWP and WLW methods can be applied to the case with the multistate failure time data. For example, in a typical AIDS study to evaluate treatment effects, the outcome measures may be the time to develop a particular type of serious infection and the time to death.[19]

4. ANALYSIS OF FAILURE TIMES FOR UNORDERED AND DISTINCT EVENTS

Consider a placebo controlled trial to evaluate the oral fluconazole, a potent antifungal agent, for the prevention of oral and vaginal candidiasis in HIV-infected women at high risk for a candida infection. Each study patient potentially has two event times, one is the time to the infection in the oral mucosa and the other is the infection time in the vaginal mucosa.

The two-sample non-parametric tests[15] discussed in Section 3 can certainly be used for the present case to test if there is a treatment different with respect to the bivariate infection times. However, since there is no natural ordering (for example, the oral infection does not always precede the vaginal infection) among these two types of event times, the AG and PWP methods cannot be applied directly to the present case. On the other hand, the WLW model and the multivariate AFT model[26] can easily handle this type of multiple event times. The frailty model discussed in the previous section can also be used in the analysis of unordered and distinct event times.

5. ANALYSIS OF CLUSTERED FAILURE TIMES

Consider the Diabetic Retinopathy Study (DRS), a randomized trial conducted by the National Eye Institute to evaluate laser photocoagulation treatment for proliferative diabetic retinopathy.[32] Between 1972 and 1975, 1742 patients were enrolled in the study. Photocoagulation was randomly assigned to one eye of each patient, with the other eye serving as an untreated control. One of the major goals of the study was to investigate if the time of occurrence of severe visual loss for the treated eye is longer than that for the control. Some potentially important covariates associated with the failure time variable, such as the presence or absence of macular oedema of the eye, and age and gender of the patient were also recorded. Note that the AG, PWP and WLW methods described in Section 3 cannot be used to analyse the correlated failure time data from DRS.

The stratified Cox procedure,[33] which allows for separate baseline hazard for each patient, can be used to evaluate the treatment effect for the DRS. However, it is not applicable to the data from studies such as the Diabetes Control and Complications Trial.[34] In that study, patients were randomly assigned to receive either experimental or standard therapy. The purpose of the study was to assess the relationship between glycaemia control and diabetic complications, including diabetic retinopathy, in persons with insulin-dependent diabetes mellitus. Experimental therapy involved the use of an intensive insulin regimen designed to maintain near normal glycaemia levels in the absence of severe hypoglycaemia. Standard treatment was designed to maintain patients free of clinical symptoms related to hyperglycaemia or hypoglycaemia while receiving up

to two insulin injections daily. Since, the treatment is systematic and affects both eyes, the Cox stratified procedure cannot be used to analyse this type of data.

Recently, Lee *et al.*[35] have used the general estimating equations approach[11] to derive an inference procedure for the regression parameters based on the above clustered failure times. They model the failure time for each eye based on the Cox model with a common nuisance hazard function between two eyes. The software for this method is available in MULCOX2, the SAS PHREG and S macros. One may also model the failure time for each eye with the AFT model. Then, using the generalized estimating equations approach, we can obtain inference procedures for the treatment difference.[36] The Cox frailty model or the usual linear mixed effects model has an intuitive physical interpretation and is a potentially useful tool for the analysis of such clustered failure times.

6. REMARKS

Methods for the analysis of censored multiple outcome data are under extensive development. The WLW[13] and other similar methods[25,35,36] are robust and well-developed. They are excellent tools for making inferences on the 'population average'[37] effect of covariates on failure times. They provide, however, no insights into the interrelationship among failure times. In some multiple outcome problems (for example, for the case with the recurrence events), patients are not only interested in the marginal covariate effects on the risk of individual failures, but also how their prior events affect the risk of having future failures. To answer such questions, the conditional approaches[16,17] with fewer modelling assumptions would be appropriate. The frailty model, which is appealing to the practitioners, can also be used to explore the relationship among distinct failures. However, in our opinion, the AFT model would be more attractive than the Cox counterpart. For the AFT mixed effects model, the fixed effects would be the same either conditional on the frailty or marginally.

REFERENCES

1. Miller, R. G. *Simultaneous Statistical Inferences*, Springer-Verlag, New York, 1977.
2. Fisher, L. D. 'A review of methods for handling multiple endpoints in clinical trials', *Proceedings of the American Statistical Association*, 43–46 (1991).
3. Cox, D. R. 'Regression models and life-tables' (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–222 (1972).
4. Lin, D. Y. 'Cox regression analysis of multivariate failure time data: the marginal approach', *Statistics in Medicine*, **13**, 2233–2247 (1994).
5. Hoppe, F. M. *Multiple Comparisons, Selection, and Applications in Biometry*, Dekker, New York, 1993.
6. Marcus, R., Peritz, E. and Gabriel, K. R. 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **63**, 655–660 (1976).
7. Holm, S. 'A simple sequential rejective multiple test procedure', *Scandinavian Journal of Statistics*, **6**, 65–70 (1979).
8. O'Brien, P. C. 'Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079–1087 (1984).
9. Wei, L. J. and Johnson, W. E. 'Combining dependent tests with incomplete repeated measurements', *Biometrika*, **72**, 359–364 (1985).
10. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
11. Liang, K. Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
12. Byar, D. and Blackard, C. 'Comparison of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage I bladder cancer', *Urology*, **10**, 556–561 (1977).

13. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065–1073 (1989).
14. Gill, R. D. *Censoring and Stochastic Integral* (*Tract 124*), Mathematical Centre, Amsterdam, 1980.
15. Wei, L. J. and Lachin, J. M. 'Two-sample asymptotically distribution-free tests for incomplete multivariate observations', *Journal of the American Statistical Association*, **79**, 653–661 (1984).
16. Andersen, P. K. and Gill, R. D. 'Cox's regression model for counting processes: a large sample study', *Annals of Statistics*, **10**, 1100–1120 (1982).
17. Prentice, R. L., Williams, B. J. and Peterson, A. V. 'On the regression analysis of multivariate failure time data', *Biometrika*, **68**, 373–379 (1981).
18. Prentice, R. L. and Cai, J. 'Marginal and conditional models for the analysis of multivariate failure time data', *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kulwer Academic Publishers, Dordrecht, 1992, pp. 393–406.
19. Li, Q. and Lagakos, S. W. 'Use of the Wei-Lin- Weissfeld method for the analysis of a recurrent and a terminating event', *Statistics in Medicine*, **16**, 000–000 (1997).
20. Lin, D. Y. 'MULCOX2: a general computer program for the Cox regression analysis of multiple failure time variables', *Computer Methods and Program in Biomedicine*, **32**, 125–135 (1990).
21. Miller, R. 'Least squares regression with censored data', *Biometrika*, **63**, 449–464 (1976).
22. Prentice, R. L. 'Linear rank tests with right-censored data', *Biometrika*, **65**, 167–179 (1978).
23. Buckley, J. and James, I. 'Linear regression with censored data', *Biometrika*, **66**, 429–436 (1979).
24. Tsiatis, A. A. 'Estimating regression parameters using linear rank tests for censored data', *Annals of Statistics*, **18**, 354–372 (1990).
25. Wei, L. J., Ying, Z. and Lin, D. Y. 'Linear regression analysis of censored survival data based on rank tests', *Biometrika*, **77**, 845–851 (1990).
26. Lin, J. S. and Wei, L. J. 'Linear regression analysis for multivariate failure time observations', *Journal of the American Statistical Association*, **87**, 1091–1097 (1992).
27. Clayton, D. G. and Cuzick, J. 'Multivariate generalizations of the proportional hazards model' (with discussion), *Journal of the Royal Statistical Society*, *Series B*, **148**, 82–117 (1985).
28. Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. 'A counting process approach to maximum likelihood estimation in frailty models', *Scandinavian Journal of Statistics*, **19**, 25–43 (1992).
29. Murphy, S. A. 'Asymptotic theory for the frailty model', Technical Report 108, Department of Statistics, Pennsylvania State University, 1992.
30. Murphy, S. A. 'Consistency in a proportional hazards model incorporating a random effect', *Annals of Statistics*, **22**, 712–731 (1995).
31. Hougaard, P. 'Survival models for heterogeneous populations derived from stable distributions', *Biometrika*, **73**, 387–396 (1986).
32. Diabetic Retinopathy Study Research Group. 'Diabetic Retinopathy Study', *Investigative Opthalmology nd Visual Science*, **21**, 149–226 (1981).
33. Holt, J. D. and Prentice, R. L. 'Survival analysis in twin studies and matched pair experiments', *Biometrika*, **61**, 17–30 (1974).
34. The DCCT Group. 'The Diabetes Control and Complications Trial: the design and methodological considerations for feasibility phase', *Diabetes*, **35**, 530–545 (1986).
35. Lee, E. W., Wei, L. J. and Amato, D. A. 'Cox-type regression analysis for large numbers of small groups of correlated failure time observations', *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kulwer Academic Publisher, Dordrecht, 1992, pp. 237–247.
36. Lee, E. W., Wei, L. J. and Ying, Z. 'Linear regression analysis for highly stratified failure time data', *Journal of the American Statistical Association*, **88**, 557–565 (1993).
37. Zeger, S. L., Liang, K. Y. and Albert, P. S. 'Models for longitudinal data: A generalized estimating equation approach', *Biometrics*, **44**, 1049–1060 (1988).

.

