STATISTICS IN MEDICINE, VOL. 13, 2233–2247 (1994)

COX REGRESSION ANALYSIS OF MULTIVARIATE FAILURE TIME DATA: THE MARGINAL APPROACH

D. Y. LIN

Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, U.S.A.

SUMMARY

Multivariate failure time data are commonly encountered in scientific investigations because each study subject may experience multiple events or because there exists clustering of subjects such that failure times within the same cluster are correlated. In this paper, I present a general methodology for analysing such data, which is analogous to that of Liang and Zeger for longitudinal data analysis. This approach formulates the marginal distributions of multivariate failure times with the familiar Cox proportional hazards models while leaving the nature of dependence among related failure times completely unspecified. The baseline hazard functions for the marginal models may be identical or different. Simple estimating equations for the regression parameters are developed which yield consistent and asymptotically normal estimators, and robust variance–covarinace estimators are constructed to account for the intra-class correlation. Simulation results demonstrate that the large-sample approximations are adequate for practical use and that ignoring the intra-class correlation could yield rather misleading variance estimators. The proposed methodology has been fully implemented in a simple computer program which also incorporates several alternative approaches. Detailed illustrations with data from four clinical or epidemiologic studies are provided.

1. INTRODUCTION

Multivariate failure time data arise when each study subject may experience several events or when there exists some natural or artificial grouping of subjects which induces dependence among failure times of the same group. Examples in biomedical research are the sequence of tumour recurrences or infection episodes, the development of physical symptoms or diseases in several organ systems, the occurrence of blindness in the left and right eyes, the onset of a genetic disease among family members, the initiation of cigarette smoking by classmates, and the appearance of tumours in littermates exposed to a carcinogen. Examples in other areas include the repeated breakdowns of a certain type of machinery in industrial reliability, the experiences of different life events by each person in sociological studies, and the purchases of various products by each consumer in marketing research. Described below are four biomedical studies which involve multivariate failure times.

Example 1.1 (The Colon Cancer Study). A national intergroup trial was conducted in the 1980's to study the drugs levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma.[1,2] In the study, 929 patients with stage C disease were randomly assigned to observation, levamisole alone, or levamisole combined with fluorouracil. The time to cancer recurrence and the survival time were both considered important outcome measures.

CCC 0277–6715/94/212233–15
Received April 1993
Revised August 1993

Example 1.2 (The CGD Study). Chronic granulomatous disease (CGD) is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections which may lead to death. In order to study the ability of gamma interferon to reduce the rate of infections, a placebo controlled randomized trial was conducted by the International CGD Cooperative Study Group in the late 1980's . Each patient had the potential to experience multiple infections. By the end of the trial, 30 of 65 placebo patients and 14 of 63 patients on gamma interferon had experienced at least one infection. Of the 30 placebo patients who experienced at least one infection, 5 experienced two, 4 others experienced three and 3 had four or more. Of the 14 gamma interferon patients with at least one infection, 4 experienced two and another had a third event. This study was described at greater length by Fleming and Harrington[3] (pp. 162–163). The data are listed in their Appendix D.2.

Example 1.3 (The Diabetic Retinopathy Study). The Diabetic Retinopathy Study was conducted by the National Eye Institute to assess the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy.[4] Between 1972 and 1975, 1742 patients entered the study. One eye of each patient was randomly selected for photocoagulation and the other eye was observed without treatment. The patients were followed over several years for the occurrence of blindness in the left and right eyes. One anticipates some dependence between a patient's two eyes.

Example 1.4 (The Schizophrenia Study). Dr. Ann E. Pulver of Johns Hopkins University has been conducting a genetic epidemiologic study of schizophrenia.[5] In the study, 487 first degree relatives (273 males, 214 females) of 93 female schizophrenic probands were enrolled. The number of relatives of a single proband ranges from 1 to 12. An important question is whether the risk of affective illness (depression or mania or both) in the relatives is associated with the age at onset of schizophrenia of the proband. Here, the times to affective illness are expected to be correlated among relatives of the same proband.

In the above examples, the scientific interests centre on the effects of covariates (for example, treatment) on the risk of failure. For univariate failure time data, such effects are studied mostly by the Cox proportional hazards model[6] and the associated partial likelihood principle.[7] The analysis of multivariate failure time data is complicated by the dependence of related failure times. With censoring, this dependence poses a greater challenge than (uncensored) longitudinal data. One useful solution that has gained increasing acceptance is the marginal hazard approach due to Wei, Lin and Weissfeld[8] and Lee, Wei and Amato[9] (hereafter referred to as WLW and LWA), in which the marginal distributions of multivariate failure times satisfy proportional hazards models and the dependence structure for related failure times is unspecified. (The method of WLW allows the baseline hazard functions to be different among the marginal models whereas the method of LWA postulates a common baseline hazard function.) As in the case of longitudinal data, [10] simple estimating equations can be constructed to yield consistent and asymptotically normal estimators for the regression parameters provided only that the marginal models correctly specified, and robust variance–convariance estimators can be obtained which properly account for the dependence.

The purpose of this paper is to provide a practical guide for applying the marginal approach. In the next section, I develop heuristically a general methodology which incorporates the results of WLW and LWA. I also compare the marginal approach with the methods of Andersen and Gill[11] and Prentice *et al.*[12] for analysing recurrence events. In Section 3, I illustrate in detail the use of the general methodology with the four examples cited above. In Section 4, I discuss a number of related issues.

2. METHODS

2.1. Univariate failure time data

We first review the basic results for the univariate case. Under the proportional hazards model,[6] the hazard function for the failure time T associated with a $p \times 1$ vector of possibly time-varying covariates $Z = (Z_1, \ldots, Z_p)'$ is

$$\lambda(t; Z) = \lambda_0(t)e^{\beta' Z(t)},$$

where β is a $p \times 1$ vector of unknown regression parameters, and $\lambda_0(t)$ is an unspecified baseline hazard function. When T is subject to right-censorship, we observe $X = \min\,(T,C)$ and $\Delta = I(T \leqslant C)$, where C is the censoring time and $I(\mathscr{A})$ indicates, by the values 1 versus 0, whether or not the event $\mathscr{A}$ occurs. Assume that T and C are independent conditional on Z. Let (X_i, Δ_i, Z_i) $(i = 1,\ldots,n)$ be n independent replicates of (X,Δ,Z). Then the partial likelihood function[7] for β is

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta' Z_i(X_i)}}{\sum_{j=1}^{n} Y_j(X_i) e^{\beta' Z_j(X_i)}} \right\}^{\Delta_i},$$

where $Y_j(t) = I(X_j \geqslant t)$. The corresponding score function $\partial \log L(\beta)/\partial\beta$ equals

$$U(\beta) = \sum_{i=1}^{n} \Delta_i \left\{ Z_i(X_i) - \frac{S^{(1)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} \right\},$$

where $S^{(0)}(\beta,t) = \sum_{j=1}^{n} Y_j(t)e^{\beta' Z_j(t)}$ and $S^{(1)}(\beta,t) = \sum_{j=1}^{n} Y_j(t)e^{\beta' Z_j(t)} Z_j(t)$. The maximum partial likelihood estimator $\hat{\beta}$ is the solution to $\{U(\beta) = 0\}$.

For large n, the score statistic $U(\beta)$ is approximately p-variate normal with mean 0 and with (estimated) convariance matrix $A(\hat{\beta})$, and $\hat{\beta}$ is approximately p-variate normal with mean β and with (estimated) covariance matrix $A^{-1}(\hat{\beta})$, where

$$A(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \Delta_i \left\{ \frac{S^{(2)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} - \frac{S^{(1)}(\beta, X_i) S^{(1)}(\beta, X_i)'}{S^{(0)}(\beta, X_i)^2} \right\},$$

and $S^{(2)}(\beta,t) = \sum_{j=1}^{n} Y_j(t)e^{\beta' Z_j(t)} Z_j(t) Z_j(t)'$. For testing $\beta = 0$, the non-parametric statistic $U'(0)A^{-1}(0)U(0)$ is known as the logrank statistic. Under misspecified Cox models, the robust variance–covariance estimator[13] for $\hat{\beta}$ is $A^{-1}(\hat{\beta})B(\hat{\beta})A^{-1}(\hat{\beta})$, where $B(\beta) = \sum_{i=1}^{n} W_i(\beta)W_i(\beta)'$ and

$$W_i(\beta) = \Delta_i \left\{ Z_i(X_i) - \frac{S^{(1)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} \right\} - \sum_{j=1}^{n} \frac{\Delta_j Y_i(X_j) e^{\beta' Z_i(X_j)}}{S^{(0)}(\beta, X_j)} \left\{ Z_i(X_j) - \frac{S^{(1)}(\beta, X_j)}{S^{(0)}(\beta, X_j)} \right\}.$$

2.2. Marginal approach for multivariate failure time data

We now consider the multivariate case. Suppose that there are n units and that each unit can potentially experience K types of failures. In Examples 1.1–1.3, each patient constitutes a unit, and in Example 1.4 the unit is the proband. In some situations (for example, Examples 1.1 and 1.2), there is a clear distinction of different failure types (so that the numbering of failure types needs to be consistent across units), while in others (for example, Examples 1.3 and 1.4), the failure types are indistinguishable (so that the ordering of failure types within a unit is arbitrary). To be more specific, cancer recurrence is very different from death in Example 1.1 whereas a left eye is biologically the same as a right eye in Example 1.3. (In the latter case, it would be more precise to say that there are K failures of the same type than K types of failures. To keep our statements concise, however, we will abuse the language.) If there are unequal numbers of failure types among the units, as in Example 1.4, we let K be the maximum number of failure types in a unit.

Let T_{ik} be the time when the kth type of failure occurs on the ith unit, and let C_{ik} be the corresponding censoring time. Define $X_{ik} = \min\,(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} \leqslant C_{ik})$. Also, let $Z_{ik} = (Z_{1ik}, \ldots, Z_{pik})'$ denote the covariate vector for the ith unit with respect to the kth type of failure. The failure time vector $T_i = (T_{i1}, \ldots, T_{iK})$ and the censoring time vector $C_i = (C_{i1} \ldots, C_{iK})$ are assumed to be independent conditional on the covariate vector $Z_i = (Z'_{i1} \ldots, Z'_{iK})\,(i = 1, \ldots, n)$. We further assume that (X_i, C_i, Z_i) $(i = 1, \ldots, n)$ are independent and identically distributed random elements. If T_{ik} or Z_{ik} is missing, we set $C_{ik} = 0$, which ensures that $X_{ik} = 0$ and $\Delta_{ik} = 0$. Naturally, such cases make no contribution to the calculation of the statistics. We require that data are missing completely at random.[14]

It is natural to formulate the marginal distribution for each type of failure with a proportional hazards model. Depending on whether the baseline hazard functions are identical or are different among the K types of failures, the hazard function of the ith unit for the kth type of failure is

$$\lambda_k(t; Z_{ik}) = \lambda_0(t) e^{\beta' Z_{ik}(t)}, \tag{1}$$

or

$$\lambda_k(t; Z_{ik}) = \lambda_{0k}(t) e^{\beta' Z_{ik}(t)}, \tag{2}$$

where $\lambda_0(t)$ and $\lambda_{0k}(t)$ $(k = 1, \ldots, K)$ are unspecified baseline hazard functions, and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters. In some applications (for example, Examples 1.1 and 1.2), it is necessary to allow $\lambda_{0k}(t)$ $(k = 1, \ldots, K)$ to be different, whereas in others (for example, Examples 1.3 and 1.4), it suffices to assume a common baseline hazard function. In both models (1) and (2), we take β to be the same among the marginal models. This entails no loss of generality since the assumption can always be achieved by introducing appropriate type-specific covariates, as elaborated in Section 3. Note that WLW considered model (2) with type-specific regression parameters whereas LWA studied model (1).

For the moment, pretend that the observations within the same unit are independent. Then the 'partial likelihood functions' for β are

$$\tilde{L}(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} \sum_{l=1}^{K} Y_{jl}(X_{ik}) e^{\beta' Z_{jl}(X_{ik})}} \right\}^{\Delta_{ik}} \tag{3}$$

under model (1) and

$$\tilde{L}(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) e^{\beta' Z_{jk}(X_{ik})}} \right\}^{\Delta_{ik}} \tag{4}$$

under model (2), where $Y_{ik}(t) = I(X_{ik} \geqslant t)$. (Notice the difference in the denominator between (3) and (4)). The corresponding 'score functions' are

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})} \right\} \tag{5}$$

and

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta, X_{ik})}{S_k^{(0)}(\beta, X_{ik})} \right\}, \tag{6}$$

where $S_k^{(0)}(\beta,t) = \sum_{j=1}^{n} Y_{jk}(t)\mathrm{e}^{\beta' Z_{jk}(t)}$, $S_k^{(1)}(\beta,t) = \sum_{j=1}^{n} Y_{jk}(t)\mathrm{e}^{\beta' Z_{jk}(t)} Z_{jk}(t)$ $(k = 1,\dots,K)$ and $\bar{S}^{(r)}(\beta,t) = \sum_{k=1}^{K} S_k^{(r)}(\beta,t)$ $(r = 0,1)$. In both cases, we obtain the unique estimator $\tilde{\beta}$ by solving $\{\tilde{U}(\beta) = 0\}$.

Although observations are generally correlated within the same unit, the estimator $\tilde{\beta}$ can be proven to be consistent for β as long as the marginal models are correctly specified. The derivative matrix $-\partial^2 \log \tilde{L}(\beta)/\partial\beta^2|_{\beta=\hat{\beta}}$, however, does not provide a valid variance–covariance estimator for $\tilde{U}(\beta)$. As shown in WLW and LWA, by approximating $\tilde{U}(\beta)$ with a sum of n independent and identically distributed random vectors, we can establish the asymptotic normality of $\tilde{U}(\beta)$ and obtain its limiting covariance matrix. Then the asymptotic distribution for $\tilde{\beta}$ follows from the Taylor series expansion. The main results are stated in the following paragraph.

For large n and relatively small K, the statistic $\tilde{U}(\beta)$ is approximately p-variate normal with mean 0 and with (estimated) covariance matrix $\tilde{B}(\tilde{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \tilde{W}_{ik}(\tilde{\beta})\tilde{W}_{il}(\tilde{\beta})'$, where under models (1) and (2), respectively,

$$\tilde{W}_{ik}(\beta) = \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})}\right\} - \sum_{j=1}^{n}\sum_{l=1}^{K} \frac{\Delta_{jl} Y_{ik}(X_{jl})\mathrm{e}^{\beta' Z_{ik}(X_{jl})}}{\bar{S}^{(0)}(\beta,X_{jl})}\left\{Z_{ik}(X_{jl}) - \frac{\bar{S}^{(1)}(\beta,X_{jl})}{\bar{S}^{(0)}(\beta,X_{jl})}\right\}$$

and

$$\tilde{W}_{ik}(\beta) = \Delta_{ik}\left\{Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})}\right\} - \sum_{j=1}^{n} \frac{\Delta_{jk} Y_{ik}(X_{jk})\mathrm{e}^{\beta' Z_{ik}(X_{jk})}}{S_k^{(0)}(\beta,X_{jk})}\left\{Z_{ik}(X_{jk}) - \frac{S_k^{(1)}(\beta,X_{jk})}{S_k^{(0)}(\beta,X_{jk})}\right\}.$$

Furthermore, the estimator $\tilde{\beta}$ is approximately p-variate normal with mean β and with (estimated) covariance matrix $\tilde{D}(\tilde{\beta}) = \tilde{A}^{-1}(\tilde{\beta})\tilde{B}(\tilde{\beta})\tilde{A}^{-1}(\tilde{\beta})$, where

$$\tilde{A}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{\frac{\bar{S}^{(2)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})} - \frac{\bar{S}^{(1)}(\beta,X_{ik})\bar{S}^{(1)}(\beta,X_{ik})'}{\bar{S}^{(0)}(\beta,X_{ik})^2}\right\}$$

under model (1) and

$$\tilde{A}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{\frac{S_k^{(2)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})} - \frac{S_k^{(1)}(\beta,X_{ik})S_k^{(1)}(\beta,X_{ik})'}{S_k^{(0)}(\beta,X_{ik})^2}\right\}$$

under model (2), $S_k^{(2)}(\beta,t) = \sum_{j=1}^{n} Y_{jk}(t)\mathrm{e}^{\beta' Z_{jk}(t)} Z_{jk}(t)Z_{jk}(t)'$ $(k = 1,\dots,K)$ and $\bar{S}^{(2)}(\beta,t) = \sum_{k=1}^{K} S_k^{(2)}(\beta,t)$.

Note that $\tilde{A}(\beta) = -\partial^2 \log \tilde{L}(\beta)/\partial\beta^2$. In the case of $K = 1$, the matrix $\tilde{D}(\tilde{\beta})$ reduces to the Lin–Wei robust variance–covariance estimator given at the end of Section 2.1. If the marginal models are correctly specified and if the observations' failure times within the same unit are independent, then $\tilde{B}(\tilde{\beta})$ is asymptotically equivalent to $\tilde{A}(\tilde{\beta})$. In the sequel, I refer to $\tilde{A}^{-1}(\tilde{\beta})$ and $\tilde{D}(\tilde{\beta})$ as, respectively, the naive and robust variance–covariance estimators for $\tilde{\beta}$, and call $\tilde{U}'(0)\tilde{A}^{-1}(0)\tilde{U}(0)$ and $\tilde{U}'(0)\tilde{B}^{-1}(0)\tilde{U}(0)$ the naive and robust logrank statistics, respectively. (A two-sample robust logrank test was previously studied by Wei and Lachin[15].) It is important to realize that the robust logrank test is always valid (that is, free of any model assumptions) since the marginal models are guaranteed to hold under $\beta = 0$.

In addition to drawing inferences about individual covariate effects, it is often of interest to test hypotheses involving several components of β. The multivariate general linear hypothesis can be expressed as H_0: $L\beta = d$, where L is a $r \times p$ matrix of constants and d is a $r \times 1$ vector of constants. The robust Wald statistic for testing H_0 is $(L\tilde{\beta} - d)'\{L\tilde{D}(\tilde{\beta})L'\}^{-1}(L\tilde{\beta} - d)$, which has an approximate χ^2 distribution with r degrees of freedom.

Table I. Summary statistics for the simulation studies*

Design	β	n	Model (1)						Model (2)					
					SEE		Size/power				SEE		Size/power	
			Bias	SSE	Na.	Ro.	Na.	Ro.	Bias	SSE	Na.	Ro.	Na.	Ro.
1	0	50	0·002	0·211	0·245	0·206	0·022	0·055	0·001	0·216	0·252	0·205	0·021	0·060
		100	0·001	0·148	0·172	0·146	0·021	0·056	0·001	0·150	0·175	0·145	0·021	0·057
		200	0·001	0·103	0·121	0·103	0·021	0·051	0·001	0·104	0·122	0·103	0·021	0·053
	0·5	50	0·008	0·209	0·240	0·203	0·575	0·707	0·005	0·215	0·246	0·203	0·541	0·697
		100	0·004	0·147	0·168	0·143	0·891	0·942	0·003	0·148	0·171	0·143	0·878	0·939
		200	0·002	0·103	0·118	0·101	0·997	0·999	0·002	0·104	0·119	0·101	0·996	0·999
2	0	50	0·000	0·285	0·249	0·275	0·081	0·055	0·000	0·287	0·252	0·275	0·079	0·058
		100	− 0·002	0·198	0·174	0·194	0·083	0·052	− 0·002	0·198	0·175	0·194	0·083	0·053
		200	− 0·001	0·139	0·122	0·137	0·085	0·054	− 0·001	0·139	0·122	0·137	0·085	0·054
	0·5	50	0·007	0·279	0·243	0·267	0·546	0·473	0·007	0·282	0·246	0·268	0·539	0·471
		100	0·003	0·192	0·170	0·189	0·823	0·762	0·002	0·193	0·171	0·189	0·818	0·759
		200	0·002	0·134	0·119	0·133	0·980	0·968	0·002	0·135	0·119	0·133	0·979	0·966

* Bias and SSE are, respectively, the sampling bias and sampling standard error of $\tilde{\beta}$. SEE is the sampling mean of the standard error estimates. Na. and Ro. stand for the naive and robust statistics, respectively. The level of significance is 0.05.

2.3. Simulation studies

Monte Carlo simulations were conducted to evaluate the aforementioned inference procedures. Paired failure times with marginal hazard rates $e^{\beta Z_{ik}}$ ($i = 1,\ldots,n$; $k = 1,2$) were generated from Gumbel's bivariate exponential distribution[16] with correlation coefficient equal to 0·25. (Only one covariate per failure type was used.) Since the failure times were generated with a common baseline hazard function, both models (1) and (2) were true. The covariate values were generated by two different designs. Under the first design, $Z_{i1} = 1$ or 0 with equal probability, and $Z_{i2} = 0$ if $Z_{i1} = 1$ and $Z_{i2} = 1$ if $Z_{i1} = 0$. Under the second design, $Z_{i1} = Z_{i2} = 1$ or 0 with equal probability. Note that the first design corresponds to the matched pairs study and the second design to the group randomization study (in which the group is the randomization unit). Under both designs, the paired failure times were censored independently by a uniform random variable on (0,3), resulting in about 30 per cent censored observations. Table I summarizes the results for the combinations of $n = 50$, 100, 200 and $\beta = 0$, 0·5. For each combination, 10,000 data sets were generated. We draw the following conclusions from Table I and related studies:

1. The bias of $\tilde{\beta}$ is negligible. There is also little bias for the robust standard error estimator at least for large n. The robust Wald (or logrank) test has proper size, though it may be slightly anti-conservative in small and moderate samples. These conclusions hold for both designs under both models (1) and (2).
2. The analysis under model (1) tends to be more efficient than that of model (2), as reflected by the sampling standard error of $\tilde{\beta}$ and by the power of the Wald test. The difference, however, is very small, especially for large n.
3. The naive variance estimator considerably overestimates the true sampling variance under the first design and seriously underestimates the true sampling variance under the second design. Consequently, the naive Wald (or logrank) test has much lower power than the robust test under the first design, and the naive test is not valid under the second design.

2.4. Marginal versus conditional approaches for recurrent events

The choice of time scales for recurrence data needs some discussion. For the marginal approach, T_{ik} is defined as the time from study entry to the kth recurrence for the ith subject ($i = 1,..,n$; $k = 1,\dots,K$). This time scale, termed total time, is particularly appealing when the recurrences are of different natures. In some applications, it is of interest to study the times between consecutive recurrences (gap times); for example, demographers sometimes study women's birth intervals. The main difficulty in analysing gap times is that the subjects who have not experienced the kth recurrence have to be excluded from the analysis of the gap time between the kth and $(k+1)$th recurrences, which violates the assumption of missing completely at random.

Andersen and Gill[11] and Prentice, Williams and Peterson[12] (hereafter referred to as AG and PWP) have suggested two alternative approaches to analysing recurrence data. Under the AG multiplicative intensity model, the risk of a recurrent event for a subject satisfies the usual proportional hazards model, and is unaffected by earlier events that occurred to the subject unless terms that capture such dependence are included explicitly in the model as covariates. PWP specified that the hazard function at time t for the kth recurrence of the ith unit, conditional on the entire failure, censoring and covariate history prior to time t in the unit, takes the form

$$\lambda_{ik}(t) = \lambda_{ok}(t)\, e^{\beta' Z_{ik}(t)} \tag{7}$$

or

$$\lambda_{ik}(t) = \lambda_{ok}(t - t_{k-1})\, e^{\beta' Z_{ik}(t)}, \tag{8}$$

where t_{k-1} is the time of the $(k-1)$th failure ($t_0 = 0$). Model (7) pertains to total times whereas model (8) uses gap times. The interpretation of the parameters in models (7) and (8) is somewhat awkward because they are conditional on the failure and censoring information. Both the AG and PWP models are analysed by the partial likelihood principle. As demonstrated by WLW, the AG and PWP procedures are sensitive to misspecification of the dependence structure.

For computational purposes, one may cast the AG and PWP methods within the general framework for the marginal approach described previously. By redefining the risk-set indicators $Y_{ik}(t)$ as $I(X_{i,k-1} < t \leqslant X_{ik})$, instead of $I(X_{ik} \geqslant t)$, with $X_{i0} = 0\,(i = 1,\dots,n;\ k = 1,\dots,K)$ (3) and (4) become the partial likelihood functions for the AG model and model (7) of PWP, respectively. The partial likelihood function for model (8) can be obtained from (4) by replacing $Y_{jk}(X_{ik})$ with $Y^*_{jk}(G_{ik})$ and $Z_{jk}(X_{ik})$ with $Z_{jk}(X_{j,k-1} + G_{ik})$, where $G_{ik} = X_{ik} - X_{i,k-1}$ and $Y^*_{jk}(t) = I(G_{jk} \geqslant t)$. In either the case of AG or that of PWP, $\tilde{A}^{-1}(\tilde{\beta})$ is the variance–covariance estimator for the resulting parameter estimator $\tilde{\beta}$.

Which of these three approaches should be used to analyse recurrent events? If one is only interested in the overall rate for recurrences of the same nature, the easiest to use seems to be the AG model (with appropriate time-dependent covariates to capture the dependence) especially when there are only a few second recurrences. If the main interest lies in gap times, then the PWP approach may be used. On the other hand, the marginal approach is the most robust for analysing total times. It is recommended that the three types of models be fit to the same data set as they provide somewhat different insights.

2.5. An alternative marginal approach for clustered data

Recently, Liang *et al.*[17] suggested a different procedure for analysing model (1). Their estimating function is similar to (5), but they replaced $\bar{S}^{(1)}/\bar{S}^{(0)}$ by an analogue which exploits pairwise comparisons of independent observations. The actual form of their estimating function is

$$\sum_{i=1}^{n} \sum_{k=1}^{K} I\{n_i(X_{ik}) > 0\}\Delta_{ik}\left\{Z_{ik}(X_{ik}) - n_i^{-1}(X_{ik}) \sum_{j \neq i} \sum_{l} \mathrm{e}_{ik,jl}(\beta, X_{ik})\right\},$$

where $n_i(t) = \sum_{j \neq i} \sum_l Y_{jl}(t)$ and

$$\mathrm{e}_{ik,jl}(\beta,t) = \frac{Y_{ik}(t)Z_{ik}(t)\,\mathrm{e}^{\beta' Z_{ik}(t)} + Y_{jl}(t)Z_{jl}(t)\,\mathrm{e}^{\beta' Z_{jl}(t)}}{Y_{ik}(t)\,\mathrm{e}^{\beta' Z_{ik}(t)} + Y_{jl}(t)\,\mathrm{e}^{\beta' Z_{jl}(t)}}.$$

The resulting estimator is consistent and asymptotically normal. It would be worthwhile to compare the efficiencies of $\tilde{\beta}$ and the Liang *et al.* estimator.

2.6. Software availability

MULCOX[18] is a FORTRAN program designed for the WLW procedures. A much more general program, called MULCOX2,[19] was developed in conjunction with the preparation of this paper. MULCOX2 implements all the methods described in Sections 2.2 and 2.4. Arbitrary patterns of time-dependent covariates and risk-set indicators are allowed. In addition, Dr. Terry Therneau of the Mayo Clinic has developed some SAS and S macros which serve similar purposes to those of MULCOX2. All these programs are available through StatLib.

3. EXAMPLES

In this section, I apply the techniques described in the last section to the four biomedical studies introduced in Section 1. For comparison, both the naive and robust statistics are presented, though the former are generally inappropriate. All the results reported in this section were obtained from MULCOX2 except for the last two columns of Table III.

3.1. The Colon Cancer Study

In this trial, 315, 310 and 304 patients with stage C disease received observation, levamisole alone, and levamisole combined with fluorouracil, respectively. Patients enrolled between March 1984 and October 1987. The study was terminated following an interim analysis in September 1989, when levamisole + fluorouracil was found to be highly effective in prolonging survival and reducing the risk of cancer recurrence. By the end of the study, 155 patients in the observation group, 144 in the levamisole alone group and 103 in the levamisole + fluorouracil group had recurrences, and there were 114, 109 and 78 deaths in the observation, levamisole alone and levamisole + fluorouracil groups, respectively. For simplicity, we focus only on the comparison between the observation and levamisole + fluorouracil (Lev + 5-FU) groups. Thus, the number of units n is 619 and the number of failure types K is 2. I treat recurrence as the first failure type and death as the second. Since recurrences occur before deaths, $\lambda_{01}(t)$ must be different from $\lambda_{02}(t)$.

Let us first consider model (2) with type-specific covariates $Z_{i1} = (R_i,0)'$ and $Z_{i2} = (0,R_i)'$ $(i = 1,\ldots,619)$, where

$$R_i = \begin{cases} 1 & \text{if the } i\text{th patient was on Lev + 5-FU,} \\ 0 & \text{if the } i\text{th patient was on observation.} \end{cases}$$

Note that $\beta' Z_{i1} = \beta_1 R_i$ and $\beta' Z_{i2} = \beta_2 R_i$ so that β_1 and β_2 pertain to the treatment effects on recurrence and death, respectively. (This illustrates that assuming a common β for the K marginal models does not preclude the use of type-specific parameters.) We are essentially fitting two separate standard Cox models to recurrence and death with the treatment indicator as the single covariate in each model, but formulation (2) permits simultaneous estimation of β_1 and β_2 as well as direct estimation of the correlation between the two estimators. We obtain $\tilde{\beta} = (-0{\cdot}517, -0{\cdot}398)'$ with naive and robust standard error estimates of $(0{\cdot}1273, 0{\cdot}1471)'$ and $(0{\cdot}1266, 0{\cdot}1475)'$, respectively. The closeness between these two sets of standard error estimates is not surprising because they are asymptotically equivalent under the current parameterization if the assumed marginal models are correct. The standardized parameter estimates (that is, estimate/standard error) based on the robust variance estimates are $(-4{\cdot}08, -2{\cdot}70)$. The naive and robust variance–covariance estimates for $\tilde{\beta}$ are

$$\tilde{A}^{-1}(\tilde{\beta}) = \begin{bmatrix} 0{\cdot}0162 & 0 \\ 0 & 0{\cdot}0216 \end{bmatrix}, \quad \tilde{D}(\tilde{\beta}) = \begin{bmatrix} 0.0160 & 0.0144 \\ 0.0144 & 0.0218 \end{bmatrix}.$$

Because of the high correlation between $\tilde{\beta}_1$ and $\tilde{\beta}_2$, the naive and robust tests for a multivariate hypothesis (involving both β_1 and β_2) can be quite different. For example, the logrank statistic for testing $\beta = 0$ is 24·27 using $\tilde{A}(\tilde{\beta})$ and 17·23 using $\tilde{B}(\tilde{\beta})$. The robust Wald statistic for testing $\beta_1 = \beta_2$ is 1.57 whereas the naive test statistic is 0.37. Apparently, there is no convincing evidence for different sizes of treatment effects on cancer recurrence and death.

We now suppose that $\beta_1 = \beta_2 = \beta$. (The null hypothesis of no treatment effect on either recurrence or death corresponds to $\beta_1 = \beta_2 = \beta = 0$. As long as β_1 and β_2 are not too far apart, the estimator of β provides a useful summary of the overall treatment difference.) By letting $Z_{i1} = Z_{i2} = R_i$ (which implies that $\beta' Z_{i1} = \beta' Z_{i2} = \beta R_i$), we obtain $\tilde{\beta} = -0{\cdot}466$ with naive standard error estimate of 0·096 and robust standard error estimate of 0·128, the corresponding standardized parameter estimates being $-4{\cdot}84$ and $-3{\cdot}65$, respectively. (Unlike the estimation of *separate* treatment effects discussed in the preceding paragraph, the naive and robust standard error estimators for the *common* treatment effect are not asymptotically equivalent if the two failure types are correlated.) The use of the robust standardized estimate or the robust logrank statistic (the latter being 13·54) for the common parameter β would enable one to make a single probability statement regarding the overall benefit of Lev + 5-FU. This would be particularly attractive if a stopping rule incorporating both outcome measures were desired.[20]

There were some imbalances between the observation and Lev + 5-FU groups with respect to certain prognostic factors. Thus, it is desirable to run a confirmatory analysis which adjusts for the prognostic variables. To this end, we fit model (2) with $Z_{i1} = Z_{i2} = (R_i, S_i, D_i, N_i)'$, where

$$S_i = \begin{cases} 1 & \text{if the surgery for the } i\text{th patient took place } \leqslant 20 \text{ days prior to randomization,} \\ 0 & \text{if the surgery for the } i\text{th patient tool place } > 20 \text{ days prior to randomization;} \end{cases}$$

$$D_i = \begin{cases} 1 & \text{if the depth of invasion for the } i\text{th patient was submucosa or muscular layer,} \\ 0 & \text{if the depth of invasion for the } i\text{th patient was serosa;} \end{cases}$$

$$N_i = \begin{cases} 1 & \text{if the number of nodes involved in the } i\text{th patient was 1–4,} \\ 0 & \text{if the number of nodes involved in the } i\text{th patient } > 4. \end{cases}$$

This analysis yields $-0{\cdot}483$ as the estimate for the common treatment effect with robust standard error estimate of 0·131. The depth of invasion and the number of nodes are both highly significant.

Eleven of the 619 patients in the observation and Lev + 5-FU groups died without cancer recurrences. In the above analyses, recurrence times on those patients were censored at deaths. Strictly speaking, the assumption of conditional independence between the failure time and the censoring time is not completely satisfied in such analyses. To avoid this problem, one may consider deaths without recurrences as events for the first failure type. Then the first failure time variable is interpreted as recurrence-free survival time, that is, time to either cancer recurrence or death. For this study, very similar results were obtained between the two approaches mainly because less than 2 per cent of the patients died without recurrences, compared to 42 per cent who had recurrences first. For the model considered in the preceding paragraph, that is, model (2) with $Z_{i1} = Z_{i2} = (R_i, S_i, D_i, N_i)'$, we obtain $\tilde{\beta}_1 = -0{\cdot}467$ with a robust standard error estimate of 0·130 when deaths without recurrences are treated as events for the first failure type.

3.2. The CGD Study

The main statistical analysis of the CGD study was based on time to the first infection. By fitting the standard Cox model with the treatment indicator R ($R_i = 1$ if the ith patient was on gamma interferon and $R_i = 0$ otherwise) as the single covariate, we obtain $\hat{\beta} = -1{\cdot}094$ with standard error estimate of 0·335. (Note that our numbers are different from those of Fleming and Harrington[3] (p. 163) because they used only the infections that had occurred by the interim analysis cut off whereas we make use of the additional data on occurrence of infections between the interim analysis cutoff and the final study visit for each patient. Appendix D.2 of Fleming and Harrington[3] contains the full data set used here.)

Since the investigators were interested in how gamma interferon reduces the *rate* of infections, it seems desirable to incorporate into analysis the additional data on recurrent events. The simplest way is to fit the AG multiplicative intensity model for all infection episodes with R as the single covariate. Under this Markov model, the estimate of treatment effect is $-1{\cdot}097$ with an estimated standard error of 0·261. This analysis assumed that the patient's risk for a new infection at a given time is not altered by the pattern of prior infections. As an attempt to accommodate the dependence of infection patterns, we add to the preceding model a time-dependent covariate, which indicates by the value 1 versus 0 whether or not the patient had an infection within the previous 60 days. The parameter estimate for this covariate is 0·712 with standard error estimate of 0·293, which is highly significant. In this semi-Markov model, the estimate for the treatment parameter becomes $-0{\cdot}989$ with standard error estimate of 0·266.

The representation of the infection history by simple time-dependent covariates may be inadequate. To avoid specifying the nature of dependence, we use the marginal approach. In this application, T_{ik} is the time from study enrolment to the kth infection for the ith patient and C_{ik} is the time from study enrolment to the final study visit for the ith patient. Since there were very few fourth infections, we will study only the first three infections. To estimate separate treatment effects for the three failure types, we fit model (2) with $Z_{i1} = (R_i, 0,0)', Z_{i2} = (0, R_i, 0)'$ and $Z_{i3} = (0, 0, R_i)'$ ($i = 1,\ldots, 128$); to estimate an overall treatment effect, we fit model (2) with $Z_{ik} = R_i$ ($i = 1,\ldots, 128$; $k = 1,2,3$). The results of these analyses are summarized in Table II. For comparison, we also display the results for the PWP approach using the same covariates as the marginal approach, as well as those of the two AG models. (The results for the AG models reported in the last paragraph were based on *all* infections whereas those of Table II are restricted to the *first three* infections only.)

Table II. Estimates of treatment effects for the CGD Study*

Methods	Infection number			
	1	2	3	1 ~ 3
Marginal	– 1·094	– 1·231	– 2·063	– 1·215
	(0·335)	(0·538)	(1·019)	(0·353)
PWP				
total time	– 1·094	– 0·151	– 1·279	– 0·859
	(0·335)	(0·566)	(1·084)	(0·280)
gap time	– 1·094	– 0·090	– 1·077	– 0·872
	(0·335)	(0·537)	(1·084)	(0·279)
AG				
Markov	—	—	—	– 1·020
	—	—	—	(0·267)
semi-Markov	—	—	—	– 0.943
	—	—	—	(0·269)

* The standard error estimates are given in parentheses

As shown in Table II, using any of the three approaches, one arrives at the conclusion that gamma interferon indeed reduces the infection are substantially. Compared to the PWP and AG methods, the marginal approach gives a somewhat larger estimate of the common treatment parameter along with a larger standard error estimate. Note that, for testing no overall treatment benefit, the marginal approach is always valid whereas the validity of PWP and AG methods depends on the correct specification of the dependence structure. It is interesting to observe that the PWP approach does not yield significant treatment effects for the second and third infections.

3.3. The Diabetic Retinopathy Study

I confine my attention to a subset of the data from the Diabetic Retinopathy Study (DRS) which was previously analysed by Huster *et al.*[21] and Liang *et al.*[17] The analysis subset is a 50 per cent sample of the high-risk patients as defined by DRS criteria ($n = 197$). By the end of the study, 54 treated eyes and 101 control eyes in this subsample had developed blindness.

In this example, each patient could potentially experience blindness in both eyes; therefore, there are two failure types with $k = 1$ and 2 denoting the left and right eyes, respectively. Since there are no biological differences between the left and right eyes, it is natural to assume a common baseline hazard function for the two failure types.

As mentioned is Section 1, the main hypothesis of interest is whether laser photocoagulation delays the occurrence of blindness. Because juvenile and adult diabetes have very different courses, it is desirable to examine how the age at onset of diabetes may affect the time of blindness. Following Huster *et al.* and Liang *et al.*, we consider model (1) with $Z_{ik} = (Z_{1ik}, Z_{2ik}, Z_{3ik})'$ ($i = 1, \ldots, 197$; $k = 1, 2$), where

$$Z_{1ik} = \begin{cases} 1 & \text{if the } k\text{th eye of the } i\text{th patient was on treatment,} \\ 0 & \text{otherwise;} \end{cases}$$

$$Z_{2ik} = \begin{cases} 1 & \text{if the } i\text{th patient had adult onset diabetes,} \\ 0 & \text{if the } i\text{th patient had juvenile onset diabetes;} \end{cases}$$

Table III. Estimates of regression parameters for the Diabetic Retinopathy Study*

Covariate	Methods			
	Naive	Robust	Liang	Huster
Treatment (Z_1)	− 0·425 (0·218)	− 0·425 (0·185)	− 0·422 (0·185)	− 0·43 (0·22)
Diabetic type (Z_2)	0·341 (0·199)	0·341 (0·196)	0·340 (0·196)	0·37 (0·20)
Interaction ($Z_1 \times Z_2$)	− 0·846 (0·351)	− 0·846 (0·304)	− 0·844 (0·303)	− 0·84 (0·35)

* The standard error estimates are given in parentheses

and $Z_{3ik} = Z_{1ik} \times Z_{2ik}$. The results of our analysis are presented in Table III along with those of Huster *et al.* and Liang *et al.*

The robust standard error estimates are appreciably smaller than the naive estimates. The treatment appears to be effective, and this effect is much stronger for adult onset diabetes than for juvenile onset diabetes. The Liang *et al.* method produces very similar parameter estimates to ours, and their standard error estimates are almost identical to our robust ones. Huster *et al.* specified a Weibull baseline hazard function for model (1). Their parameter estimates are fairly close to $\tilde{\beta}$ whereas their standard error estimates are similar to the naive estimates.

3.4. The Schizophrenia Study

In this ongoing genetic epidemiologic study, the failure time is the age at diagnosis of affective illness for the relative. There are only 31 events out of the 487 relatives in the current database. The covariate of major interest, the proband's age, has been dichotomized at 16 years. The gender of the relative is also expected to be predictive. We assume that gender is the only characteristic that differentiates relatives of the same proband. It is then natural to consider model (1) with $Z_{ik} = (Z_{1ik}, Z_{2ik})'$ $(i = 1,\ldots, 93;\ k = 1,\ldots, 12)$, where

$$Z_{1ik} = \begin{cases} 1 & \text{is the age at onset of the } i\text{th proband } \leqslant 16, \\ 0 & \text{otherwise;} \end{cases}$$

and

$$Z_{2ik} = \begin{cases} 1 & \text{if the } k\text{th relative to the } i\text{th proband is male,} \\ 0 & \text{if the } k\text{th relative of the } i\text{th proband is female.} \end{cases}$$

Note that we set $K = 12$, the maximal number of relatives for a proband. Since the ordering of relatives within a family is arbitrary, we suppose that the missing components occupy the tail portion of the failure vector $T_i = (T_{i1},\ldots,T_{i,12})'$ for each i. We obtain $\tilde{\beta} = (-0{\cdot}238,\ -1{\cdot}244)'$ with naive and robust standard error estimates of $(0{\cdot}489, 0{\cdot}411)'$ and $(0{\cdot}517, 0{\cdot}408)'$, respectively. Therefore, the proband's age at onset is not significant whereas the relative's gender is. The failure to establish an association between the familial risk and the proband's age at onset may be due to the small number of events. We intend to re-analyse the data after further follow-up.

4. DISCUSSIONS

Estimating functions (5) and (6) were derived under the independence working assumption. As in the case of longitudinal data,[10] it may be more efficient to use estimating functions that take into account the nature of dependence explicitly. This amounts to forming certain linear combinations of the contributions to (5) or (6) from the K types of failures. The resulting estimators remain consistent and asymptotically normal with estimable covariance matrices under mild regularity conditions on the weight matrices. Because of the censoring and the non-linear nature of the Cox model, however, it is difficult to construct optimal weight matrices. In her unpublished Ph.D. dissertation, J. Cai suggested the use of the inverse matrix of the covariance functions between counting process martingales[22] for model (2). Her simulations, however, indicated that the efficiency improvements for the resulting estimators are small unless the correlations of failure times are unusually high. For estimating a common regression parameter, WLW used a linear combination of type-specific parameter estimators which achieves the smallest asymptotic variance among all linear combinations. (For the CGD study, the WLW method estimates the overall treatment effect at $-1{\cdot}103$ with standard error estimate of $0{\cdot}333$.) In a similar spirit, Lin[20] proposed a weighted sum of the marginal logrank statistics which maximizes asymptotic power against certain local alternatives. Further research into dependent working models is warranted.

Lin[20] also addressed the issue of sequential testing in randomized clinical trials with multiple endpoints. He derived the joint distribution of the weighted sums of the marginal logrank statistics calculated at successive interim analyses. The knowledge of this joint distribution enables one to construct the stopping rule which preserves a single preset overall significance level. This approach is most appealing when the endpoints being combined exhibit treatment differences in the same direction. It can lead to substantial savings in sample size compared to sequential methods based on single endpoints.

It is important to assess the adequacy of the marginal models. A simple method for examining the key proportional hazards assumption is to test for the significance of interaction terms between covariates and t or $\log t$, as was originally suggested by Cox[6]. One may also compare parameter estimators with different weight matrices.[23] Recently, elaborate techniques for checking (univariate) survival models have been developed by using martingale-based residuals.[24–26] Generalizations of these methods to the multivariate setting are currently being investigated by C. Spiekerman in his dissertation.

One frequent goal in fitting a survival model is to estimate the survival distribution. Analogous to the Breslow[27] estimator for the univariate Cox model, it is natural to estimate the cumulative baseline hazard functions in models (1) and (2) by $\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(X_{ik} \leqslant t)\Delta_{ik}/\bar{S}^{(0)}(\tilde{\beta}, X_{ik})$ and $\hat{\Lambda}_{0k}(t) = \sum_{i=1}^{n} I(X_{ik} \leqslant t)\Delta_{ik}/S_k^{(0)}(\tilde{\beta}, X_{ik})$, respectively. The corresponding estimators for the survival function of a subject with the set of covariate values z are $\exp\{-\hat{\Lambda}_0(t)e^{\tilde{\beta}'z}\}$ and $\exp\{-\hat{\Lambda}_{0k}(t)e^{\tilde{\beta}'z}\}$, respectively. These estimators can be shown to be consistent and asymptotically normal.

In many applications, failure times are broadly grouped. Most commonly, they arise when the (continuous) failure time is subject to interval grouping. In other instances, the time measurement may truly be discrete, as, for example, when the time represents the number of attempts required to successfully perform a certain task. For the univariate failure time variable, Prentice and Gloeckler[28] studied a grouped data version of the Cox proportional hazards model. Recently, Guo and Lin[29] extended the work of Prentice and Gloeckler to the multivariate setting. Their procedures are essentially the discrete versions of those described in Section 2.2.

A useful alternative to the proportional hazards model is the accelerated failure time model, which relates the logarithm of the failure time linearly to the covariates. Semiparametric inference for this model received considerable attention in the last few years.[30, 31] Recently, Lin and Wei[32] and Lee *et al.*[33] applied the ideas of WLW and LWA, respectively, to the case of accelerated failure time models.

The marginal approach exploited in this paper treats the dependence of related failure times as a nuisance. In contrast, a number of authors[34-37] have studied the so-called frailty models, which explicitly formulate the nature of dependence. To be specific, the hazard function for the ith unit with respect to the kth type of failure, given the frailty Q_i, takes the form

$$\lambda_{ik}(t; Z_{ik}, Q_i) = Q_i \lambda_0(t) e^{\beta' Z_{ik}(t)}, \tag{9}$$

where the frailty variables Q_i ($i = 1, \dots, n$) are postulated to follow a given parametric distribution. Conditional on Q_i ($i = 1, \dots, n$), the failure times are assumed to be independent. Note that β in (9) generally needs to be interpreted conditionally on the unobservable frailty. There has been considerable controversy over whether the unconditional specification of the marginal hazard approach or the conditional specification of the frailty model approach is more naturally related to the underlying mechanisms. The latter approach is expected to be more efficient than the former provided that the frailty distribution is correctly specified. However, the types of dependence encompassed by the frailty models are quite limited, and the model fitting is rather cumbersome. So far there has not been a general large-sample theory for frailty models, though significant progress is being made. The interested reader is referred to the recent text of Andersen *et al.*[38] for an excellent exposition of frailty models.

ACKNOWLEDGEMENTS

The author is grateful to Professors Thomas Fleming, David Harrington, Steven Self and Kung-Yee Liang for providing the data files of the Colon Cancer, CGD, DRS and Schizophrenia Studies, respectively. He would also like to thank Professors L. J. Wei and Zhiliang Ying for many stimulating discussions on the subject over the years. Suggestions from two referees led to a clearer presentation. This research was supported by the National Institutes of Health Grants R29 GM-47845 and R01 AI-291968.

REFERENCES

1. Moertel, C. G., Fleming, T. R., McDonald, J. S., MacDonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H., Veeder, M. H. and Mailliard, J. A. 'Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma', *New England Journal of Medicine*, **322**, 352–358 (1990).
2. Fleming, T. R. 'Evaluating therapeutic interventions: some issues and experiences (with discussion)', *Statistical Science*, 7, 428–456 (1992).
3. Fleming T. R. and Harrington D. P. *Counting Processes and Survival Analysis*, Wiley, New York, 1991.
4. Diabetic Retinopathy Study Research Group. 'Diabetic retinopathy study', *Investigative Ophthalmology and Visual Science*, **21**, Part 2, 149–226 (1981).
5. Pulver, A. E. and Liang, K. -Y. 'Estimating effects of proband characteristics on familial risk: II. the association between age at onset and familial risk in the Maryland schizophrenia sample', *Genetic Epidemiology*, **8**, 339–350 (1991).
6. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
7. Cox, D. R. 'Partial likelihood', *Biometrika*, **62**, 269–276 (1975).
8. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the America Statistical Association*, **84**, 1065–1073 (1989).
9. 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*, Kluwer Academic Publishers, Dordrecht, 1992, pp. 237–247.

10. Liang, K. -Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
11. 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).
12. Prentice, R. L., Williams, B. J. and Peterson, A. V. 'On the regression analysis of multivariate failure time data', *Biometrika*, **68**, 373–379 (1981).
13. Lin, D. Y. and Wei, L. J. 'The robust inference for the Cox proportional hazards model', *Journal of the American Statistical Association*, **84**, 1074–1078 (1989).
14. Rubin, D. B. 'Inference and missing values', *Biometrika*, **63**, 81–92 (1976).
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. Gumbel, E. J. 'Bivariate exponential distributions', *Journal of the American Statistical Association*, **55**, 698–707 (1960).
17. Liang, K. -Y., Self, S. G. and Chang, Y. -C. 'Modelling marginal hazards in multivariate failure time data', *Journal of the Royal Statistical Society, Series B*, **55**, 441–453 (1993).
18. Lin, D. Y. 'MULCOX: a computer program for the Cox regression analysis of multiple failure time variables', *Computer Methods and Programs in Biomedicine*, **32**, 125–135 (1990).
19. Lin, D. Y. 'MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data', *Computer Methods and Programs in Biomedicine*, **40**, 279–293 (1993).
20. Lin, D. Y. 'Nonparametric sequential testing in clinical trials with incomplete multivariate observations', *Biometrika*, **78**, 123–131 (1991).
21. Huster, W. J., Brookmeyer, R. and Self, S. G. 'Modelling paired survival data with covariates', *Biometrics*, **45**, 145–156 (1989).
22. Prentice, R. L. and Cai, J. 'Covariance and survival function estimation using censored multivariate failure time data', *Biometrika*, **79**, 495–512 (1993).
23. Lin, D. Y. 'Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators', *Journal of the American Statistical Association*, **86**, 725–728 (1991).
24. Barlow, W. E. and Prentice, R. L. 'Residuals for relative risk regression', *Biometrika*, **75**, 65–74 (1988).
25. Therneau, T. M., Grambsch, P. M. and Fleming, T. R. 'Martingale-based residuals for survival models', *Biometrika*, **77**, 147–160 (1990).
26. Lin, D. Y., Wei, L. J. and Ying, Z. 'Checking the Cox model with cumulative sums of martingale-based residuals', *Biometrika*, **80**, 573–581 (1993).
27. Breslow, N. 'Contribution to the discussion of the paper by D. R. Cox', *Journal of the Royal Statistical Society, Series B*, **34**, 216–217 (1972).
28. Prentice, R. L. and Gloeckler, L. A. 'Regression analysis of grouped survival data with applications to breast cancer data', *Biometrics*, **34**, 57–67 (1978).
29. Guo, S. W. and Lin, D. Y. 'Regression analysis of multivariate grouped survival data', *Biometrics*, in press (1994).
30. Tsiatis, A. A. 'Estimating regression parameters using linear rank tests for censored data', *Annals of Statistics*, **18**, 354–372 (1990).
31. 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).
32. Lin, J. S. and Wei, L. J. 'Linear regression analysis for multivariate failure time observations', *Journal of the American Statistical Association*, **87**, 1071–1097 (1992).
33. 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).
34. Clayton, D. and Cuzick, J. 'Multivariate generalizations of the proportional hazards model', *Journal of the Royal Statistical Society, Series A*, **148**, Part 2, 82–117 (1985).
35. Hougaard, P. 'Modelling multivariate survival', *Scandinavian Journal of Statistics*, **14**, 291–304 (1987).
36. Oakes, D. 'Bivariate survival models induced by frailties', *Journal of the American Statistical Association*, **84**, 487–493 (1989).
37. 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).
38. Andersen, P. K., Borgan, Ø, Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.