rhDNase AS AN EXAMPLE OF RECURRENT EVENT ANALYSIS

TERRY M. THERNEAU[1] AND SCOTT A. HAMILTON[2]*

[1]*Mayo Clinic, Rochester, MI, U.S.A.*
[2]*Genentech, Inc., 460 Point San Bruno Boulevard, So. San Francisco, CA 94080-4990, U.S.A.*

SUMMARY

We consider counting process methods for analysing time-to-event data with multiple or recurrent outcomes, using the models developed by Anderson and Gill, Wei, Lin and Weissfeld and Prentice, Williams and Peterson. We compare the methods, and show how to implement them using popular statistical software programs. By analysing three data sets, we illustrate the strengths and pitfalls of each method. The first example is simulated and involves the effect of a hidden covariate. The second is based on a trial of gamma interferon, and behaves remarkably like the first. The third and most interesting example involves both multiple events and discontinuous intervals at risk, and the three approaches give dissimilar answers. We recommend the AG and marginal models for the analysis of this type of data. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2029–2047 (1997)
No. of Figures: 1 No. of Tables: 9 No. of References: 23

1. INTRODUCTION

There is increasing interest, and need, to apply survival analysis to data sets with multiple events per subject. This includes both the cases of multiple events of the same type and events of different types. Examples of the former are recurrent infections in AIDS patients or multiple infarcts in a coronary study. Examples of the latter are the use of both survival and recurrence information in cancer trials, or multiple sequelae (toxicity, worsening symptoms, etc.) in the management of chronic disease. With the increasing emphasis on quality of life, rehospitalization and other secondary endpoints, such analyses will become more common.

A major issue in extending proportional hazards regression models to this situation is intra-subject correlation. Other concerns are multiple time scales, discontinuous intervals of risk, strata by covariate interactions, and the structure of the risk sets. Several approaches for dealing with such data have appeared in the literature:

(i) A counting process approach, usually called the Anderson–Gill (AG) model.[1] Each subject is treated as a multi-event counting process with essentially independent increments. The observed increments (that is, times between successive events) must be conditionally independent given the history of all observables up to the event times. One models interrelation between events as one or more time-dependent covariates. This approach is simple, but the assumptions are strong and may be untenable.

*Correspondence to: Scott A. Hamilton, Genentech, Inc., 460 Point San Bruno Boulevard, So. San Francisco, CA 94080-4990, U.S.A.

CCC 0277–6715/97/182029–19$17.50

Received April 1996
Revised November 1996

(ii) A marginal method, in which one determines $\hat{\beta}$ from a fit to all of the data, ignoring correlation, followed by calculation of a consistent estimate of the variance of $\hat{\beta}$. One such approach is developed by Wei *et al.*[2] They show the utility of the method for both a data set with multiple dissimilar outcomes (recurrence of cancer and death) and another with multiple similar outcomes (recurrence of bladder cancer).

(iii) A conditional method, such as that described in Prentice *et al.*[3] and in Oakes[4] using a frailty model. Oakes illustrates this approach using a data set from the MDPIT trial of Diltiazem; the main outcomes of interest are cardiac events. Use of the second and subsequent events gave a 10 per cent reduction in the variance of the treatment effect.

(iv) A more ambitious plan is to model the subject's correlation directly within the Cox framework. Prentice and Cai[5] explore this for a sample of industrial failure data. The method is very computer intensive, however, and, as pointed out by the discussant of their paper, required the estimation of 226 parameters from only 20 pairs of data.

In this paper we focus on the AG and marginal models. This is partly due to the availability of software for this approach in both the S-plus and SAS packages. Also, the method affords great flexibility in the formation of strata and risk sets, manipulation of the time scale, and has a well developed estimator of variance. In Section 2 we introduce the basic ideas of the marginal method, along with an important relationship between its variance estimate and the grouped jack-knife. In Section 3 we highlight issues behind choosing and setting up the AG, marginal and conditional methods. In Section 4, we discuss three examples. The first is based on simulated data and highlights some serious issues that can arise when there are hidden covariates. The second example involves treatment for chronic granulomatous disease (CGD), a chronic disease of certain immunocompromised children. Its results parallel the simulation example in several ways. The final example is the most interesting. Several aspects of the rhDNase study require serious thought: from 0–5 events per subject; intervals without risk; an apparent treatment by time interaction; and model assumptions. When applied to these data, different models give apparently different answers.

2. BACKGROUND

2.1. Cox model

For these models the most straightforward mathematical notation derives from the theory of counting processes; see Fleming and Harrington[6] or Andersen *et al.*[1] for detailed exposition. Let $Z_{ij}(t)$ be the jth covariate of the ith person (possibly time dependent). Define $r_i(t)$ as $\exp[\beta' Z_i(t)]$, that is, the risk score for the ith subject. In actual practice we replace β with $\hat{\beta}$ and the subject risks r_i with $\hat{r}_i$. For each subject we also observe two processes; the counting process $N_i(t)$ is the cumulative number of events observed for the subject; and the risk indicator $Y_i(t)$ which is 1 when the subject is at risk and under observation and 0 otherwise.

Appendix I contains detailed formulae for the standard Cox model expressions in this notation. We need only the definition of two residuals. The martingale residual, $\hat{M}_i$, is the difference between the observed and expected number of events (under the fitted model) for each subject. As such it automatically accounts for differential amounts of follow-up time. It is defined as

$$\hat{M}_i = N_i(t) - \int_0^t \hat{r}_i(s) Y_i(s) \, \mathrm{d}\hat{\Lambda}_0(s)$$

where $\hat{\Lambda}_0$ is the usual Nelson–Aalen estimate of the cumulative baseline hazard.

The most important residual for our work is the matrix D of leverage residuals, referred to as the *dfbeta* residuals in SAS or S-plus. Let L be the matrix of score residuals

$$L_{ij} = \int_0^t [Z_{ij}(s) - \bar{Z}_j(s)] \mathrm{d}\hat{M}_i(s)$$

and $D = L\mathscr{I}^{-1}$, the score residuals scaled by the variance matrix of $\hat{\beta}$. Therneau *et al.*[7] discuss the score residual, which is equivalent to the residual defined as $\hat{e}_i\{\hat{x}_i - \hat{E}(\hat{\beta}, t)]$ in Barlow and Prentice.[8]

The elements of D are the leverage residuals derived by Cain and Lange,[9] and by Reid and Crépeau[10] using another method. The ij element D_{ij} is an estimate of the change in $\hat{\beta}_j$ if we removed observation i from the sample. It is straightforward to show that the column sums $1'D$ are the Newton–Raphson increment $\Delta\hat{\beta}$ at each iteration, and thus that $1'D = 0$ at the final solution.

2.2. Computation

Computation for our models is facilitated by use of the *counting process* style of input, an option supported by the S-plus package, and more recently by SAS. The input data set consists of observations or rows of data, each of which contains (fixed) covariate values Z, a status indicator $1 =$ event, $0 =$ censored, and an optional stratum indicator, along with the time interval (start, stop] over which the information applies. In the notation above, this means that we treat each row as a separate subject whose Y_i variable is 1 on the interval (start, stop] and 0 otherwise. Within the program, it means that the risk set at time t only uses the applicable rows of data. Note that the interval is open on the left and closed on the right; this facilitates representation of a given subject as a set of intervals (observations) $(0, t_1]$, $(t_1, t_2]$,... without double counting him/her in the risk set at t_i. A status indicator of 1 indicates an event at the right hand endpoint.

This rather simple program addition allows for several extensions to the basic Cox model:

(i) Multiple events. Assume a subject has an event on days 100 and 185 and has now been followed to day 250. He would be coded as 3 observations or 'lines' of data whose intervals are (0, 100], (100, 185], and (185, 250] with corresponding status codes of 1, 1 and 0.

(ii) Time dependent covariates. The most common type of time dependent covariates are repeated measurements on a subject or a change in the subject's treatment. Both of these are straightforward in the proposed formulation. As an example, consider the well known Stanford heart transplant study,[11] where treatment is a time dependent covariate. Select two patients whose times from enrolment to death are 102 and 343 days, respectively; the second patient had a transplant 21 days from enrolment. The data file for these two would be:

Interval	Status	Transplant	Age	Prior surgery
(0, 102]	1	0	41	0
(0, 21]	0	0	48	1
(21, 343]	1	1	48	1

Note that static covariates such as age are simply repeated for a patient with multiple lines of data.

(iii) Discontinuous intervals of risk. The time intervals for a patient need not be contiguous. In a study of recurrent hip fracture, for instance, patients were not considered at risk for further fracture during hospitalization. A subject who fractured at day 100, followed by a 15 day hospital stay and then 300 more days of uneventful follow-up would be represented as two intervals:

Interval	Status
(0, 100]	1
(115, 415]	0

(iv) Other time scales. The usual Cox model forms risk groups based on time since entry. For some studies a more logical grouping might be based on another alignment, such as age or time since diagnosis. Since the interval(s) for a patient need not start at zero, these time scales are easily accommodated. Consider a subject diagnosed 1/85 who enters a study on 1/90 and is followed to 1/93. The data for this subject (in years) is (5, 8], and they are not computed in the risk sets at years 0–4, so that bias is not introduced.

(v) Time dependent strata. The strata variable, like the covariates, may also vary from observation to observation.

With these methods for setting up the data to accommodate the *counting process* style of input, the class of models that can be fitted becomes richer. It is important to note that the responsibility and difficulty for checking model assumptions both increase, and that some of those very flexible models might be very difficult to justify except in large data sets.

2.3. Robust Variance

If one suspected that some element of the Cox model were misspecified, a natural correction is to use the jack-knife estimate of variance $(J-\bar{J})'(J-\bar{J})$, where J_{ij} is the change in $\hat{\beta}_j$ with observation i removed from the data set and $\bar{J}=11'J/n$ is a matrix containing the column means of J. A natural approximation to the jack-knife, variance is $D'D$.

We can also use D to approximate a grouped jack-knife, for example, the sum of rows 1–3 of D approximates the change in $\hat{\beta}$ if we removed observations 1–3 from the data set. (This estimate is obviously cruder than for a single subject, with respect to pairs of outliers for instance). In particular, assume that the sample was formed from m groups of possibly correlated observations with

$$n=\sum_{k=1}^{m} n_k.$$

Then, one might form the collapsed $m \times p$ leverage matrix $\tilde{D}$, where

$$\tilde{D}_{1j}=\sum_{i=1}^{n_1} D_{ij}$$

$$\tilde{D}_{2j}=\sum_{i=1+n_1}^{n1+n2} D_{ij}$$

$$\vdots$$

Each row k of $\tilde{D}$ is an estimate of the leverage of the kth group, and $\tilde{D}'\tilde{D}$ approximates the grouped jack-knife estimate of variance. Hence, $D'D$ should be an asymptotically unbiased estimate of the variance of $\hat{\beta}$. The proof would parallel the many models already worked out by Lin, Wei, and others. The common use of this estimate in our work arises when there are multiple observations per subject. In this case the rows of D represent *per observation* influence and those of $\tilde{D}$ the *per subject* influence. Plots of both of these are useful in their own right for checking a fitted model.

These estimates are familiar from other contexts. Using the same method of derivation as Cain and Lange,[9] the results for a linear model are $L_{ij} = X_{ij}(y_i - \hat{y}_i)$, $D = L(X'X)^{-1}$ and $D'D$ is the robust variance estimate of White.[12,13] For a generalized linear model with log-likelihood function $l(\beta)$

$$L_{ij} = \frac{\partial l}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

and $\tilde{D}'\tilde{D}$ is the *working independence* estimate of variance proposed by Liang and Zeger[14] for generalized estimating equation (GEE) models.

Reid and Crépeau[10] mention the robust variance estimator $D'D$ briefly, but do not expand on its use. Lin and Wei[15] derive $D'D$ from another approach, and show that the estimate is robust against certain failures in the assumptions of the proportional hazards model. Wei *et al.*[2] derive and apply $\tilde{D}'\tilde{D}$ in the context of multiple event types, and show that it provides an unbiased estimate of variance in this case. Lee *et al.*[16] develop $\tilde{D}'\tilde{D}$ for correlated events of the same type.

None of the latter authors makes the connection between their estimator and the dfbeta residuals D. This is unfortunate since it gives the impression that the variance estimators are difficult to calculate, when, in fact, they obtain readily with use of any software package that returns dfbeta residuals.

3. SETTING UP THE PROBLEM

One aspect of multiple event data sets is that one has a number of choices in setting up the model. These include the choice of strata and membership within strata, time scales within strata, constructed time-dependent covariates, strata by covariate interactions, and data organization. For a 'standard' Cox model these issues are fairly well understood:

(i) Stratification, if used, is based on external variables such as enrolling institution or disease subtype. These generally correspond to predictors for which we desire a flexible adjustment, but not an estimate of the covariate effect. Each subject is in exactly one strata.

(ii) The time scale is almost invariably time since entry to the study.

(iii) Time dependent covariates usually reflect time dependent data, such as repeated laboratory tests. Strata by covariate interactions, that is, separate coefficients within each strata for some covariate, are occasionally used.

(iv) The counting process form may be used for a time dependent covariate, but normally the data set consists of one observation per subject.

In a multiple events data set there are possible extensions in each of these four areas.

The first issue is to distinguish between data sets where the multiple events have a distinct ordering and those where they do not. An example of the first is multiple sequential infections. An example of the second is the times to death and progression for a set of cancer patients. For unordered outcomes data setup is usually straightforward – each outcome is coded as a single observation, there are multiple observations per subject, and each subject has the same number of observations. Often, the analysis is stratified by observation type, for example, we assume that the baseline hazard functions for time-to-death and time-to-progression may differ. In the competing risks case where each subject may have at most one event, there is some empirical evidence that one can still use the usual variance estimator despite the correlation, see Lunn and McNeil.[17] The authors also compare models that stratify on the event type to ones that use event type as a covariate.

For ordered outcomes, that is, multiple events of the same type, there have been several suggestions offered. The most common approaches are the independent increment, marginal, or conditional models. All these are 'marginal' regression models in that we determine $\hat{\beta}$ from a fit that ignores the correlation followed by a corrected variance $\tilde{D}'\tilde{D}$, but they differ considerably in their creation of the risk sets.

3.1. Independent Increment model

This method, often referred to as the 'Andersen–Gill' formulation, is the simplest to visualize and set up, but makes the strongest assumptions. It is closest in spirit to Poisson regression, and is in fact accurately approximated with Poisson regression software in the same manner that Laird and Olivier[18] approximate an ordinary single event Cox model.

Using the counting process style of data input, each subject is represented as a set of rows with time intervals of (entry time, first event], (first event, second event], ..., (mth event, last follow-up]. A subject with 0 events would have a single observation, one with 1 event would have one or two observations (depending on whether there was additional follow-up experience after the first event), etc. Depending on the time scale, the first observation may or may not begin at zero. One alternative time scale, corresponding to a renewal process, is 'time since entry or last event' and has intervals of $(0, t_1]$, $(0, t_2 - t_1]$, ...

No extra strata or strata by covariate interaction terms are induced by the multiple events. Strata, if used, are based on the same considerations as for an ordinary single event model.

The key assumption of the model is that of independent increments, that is, that the multiple observations for a given subject are independent. Data without the independent increment structure may not satisfy the proportional hazards assumption for such a misspecified model. If so, the limiting value, β^*, of $\hat{\beta}$ may be difficult to interpret. Lin and Wei[15] have addressed this issue. For us, the important point is that in real data the proportional hazards assumption is never true, yet $\hat{\beta}$ from a Cox model has proven to be a useful statistic. Therefore, an honest estimate of its variance is desirable.

If the independent increment assumption holds, the three variance estimates $\mathscr{I}^{-1}$, $D'D$ and $\tilde{D}'\tilde{D}$ should all estimate the same quantity. Use of the clustered estimate $\tilde{D}'\tilde{D}$ allows extension of the model to data sets without independent increments. Effects that may change with event number are modelled with the time dependent covariates. For instance, let z be the time dependent covariate 'number of prior events'. A model might include both treatment, z and their interaction.

3.2. Marginal model

Wei *et al.*[2] used the marginal data model in their analysis of bladder cancer data, sometimes referred to as the 'WLW' method. For this method, each event or event type is modelled as a separate strata. Within each strata, the data used are the marginal data, that is, 'what would result if the data recorder ignored all information except the given event type'. As a result, each patient normally appears in all of the strata, barring deletion due to missing values. Usually, all the time intervals start at zero and one can fit the model without recourse to the counting process style of input.

In the WLW paper, they include all strata by covariate interaction terms in the model. In this case we can obtain the individual coefficients (and they were obtained) by fitting each stratum as a separate data set. The combined coefficient vector was then the concatenation $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ from the four fits and they estimated the combined variance as

$$\begin{pmatrix} D'_1D_1 & D'_1D_2 & D'_1D_3 & D'_1D_4 \\ D'_2D_1 & D'_2D_2 & D'_2D_3 & D'_2D_4 \\ D'_3D_1 & D'_3D_2 & D'_3D_3 & D'_3D_4 \\ D'_4D_1 & D'_4D_2 & D'_4D_3 & D'_4D_4 \end{pmatrix}$$

where D_1 is the matrix of dfbeta residuals from the first fit, D_2 that from the second etc. This is algebraically equivalent to $\hat{\beta}$ and $\tilde{D}'\tilde{D}$ from a combined fit over all four strata, where the combined model includes all covariate by strata interaction terms. Using a single fit on the combined data set is both easier to set up and more flexible, since the user is not forced to include covariate by stratum interaction terms.

3.3. Conditional model

As with the marginal model, we assign each event or event type to a different strata. The difference between the conditional and marginal models occurs when there are multiple events of the same type. For instance, assume that a patient had non-fatal myocardial infarctions on days 100 and 185, and has now been followed to day 250. In the marginal analysis this subject is at risk in stratum 2 from time 0 to 185. For the conditional model, the assumption is made that a subject cannot be at risk for event 2 until event 1 occurs; in stratum 2 this subject is at risk from time 100 to 185. Oakes[4] argues persuasively for the conditional approach, and states that the marginal method is inefficient.

3.4. Sample data

Consider subject 204001 from the CGD data in Section 4.2. This subject had infections on day 219 and 373 with further follow-up to day 414. Let the variables `t1`, `t2` mark the start and end of each interval, `status` be 1 for an event and 0 for censoring, and `enum` be a constructed variable containing the event number. For the AG and conditional models, the data set contains 3 lines for this subject:

id	t1	t2	status	enum	x1 ...
204001	0	219	1	1	···
204001	219	373	1	2	
204001	373	414	0	3	

We must construct the data set for the WLW model differently. The maximum number of infections in the data set is 7, and each subject appears in all seven strata. There is only a single time variable, and the rows for this subject are:

id	time	status	enum	x1 …
204001	219	1	1	…
204001	373	1	2	
204001	414	0	3	
204001	414	0	4	
204001	414	0	5	
204001	414	0	6	
204001	414	0	7	

Appendix II shows the SAS and S-plus code to fit the CGD examples.

4. EXAMPLES

4.1. Hidden covariate

We first illustrate the methods with a simple test case that shows each method has potential biases. Let the time to next event be exponential with rate $\exp(x_1 - x_2)$, where x_1 is uniformly distributed between -2 and $+2$ and x_2 is a randomly assigned 0/1 treatment covariate independent of x_1. Sequential events were independent. The follow-up time for each subject was 4 years, which gave a mean number of events of 3·6. The sample size was 500, which allows us to illustrate any biases in the estimate. For simplicity of presentation, we censored the few subjects with more than 10 events after their tenth. The number of events experienced was

	Number of events										
	0	1	2	3	4	5	6	7	8	9	10
Control	8	34	35	31	28	30	24	13	15	8	24
Treatment	68	64	49	33	16	10	5	2	2	0	1

First, consider the estimation of the parameters. An important point of comparison is when x_1 is *not* included in the model. This corresponds, in real data sets, to those important covariates unmeasured or unknown to us. (We purposely chose the unmeasured covariate x_1 to have a larger effect than the intervention.) In Table I we see that the time-to-first-event model underestimates β when the hidden covariate is not included. Omori and Johnson[19] have computed the amount of attenuation in the general case which is related to the variance of $\exp(x_1)$.

For the independent increments or Andersen–Gill model, Table I shows that the value of $\hat{\beta}$ is intermediate between a time to first event analysis and the true value of -1. The naive estimate of standard error is seriously underestimated when the hidden covariate is not included. When x_1 is

Table I. Simple models

	Without covariate			With covariate		
	$\hat{\beta}$	naive SE	robust SE	$\hat{\beta}$	naive SE	robust SE
First event: β_2	− 0·89	0·101	0·099	− 1·05	0·104	0·101
β_1				0·97	0·093	0·093
AG: β_2	− 0·95	0·056	0·076	− 1·00	0·056	0·060
β_1				1·05	0·051	0·047
Marginal: β_2	− 1·38	0·057	0·108	− 1·83	0·060	0·116
β_1				1·84	0·056	0·098
Conditional: β_2	− 0·69	0·061	0·066	− 1·00	0·064	0·071
β_1				1·04	0·059	0·053

included then the AG model is correctly specified, and both the coefficients and their standard errors are estimated without bias. One can show algebraically that if all subjects are followed for the same amount of time, then the AG model without the hidden covariate will correctly estimate the overall hazard ratio for these data. In this example only a few subjects, those with > 10 events, were truncated. If follow-up is truncated after 7 events the AG estimate is $-0{\cdot}91$ and quickly approaches the time to first event value of $-0{\cdot}89$ for any further truncation.

When x_1 is known the conditional model is also correctly specified, and has unbiased estimates. When x_1 is unknown the conditional model seriously underestimates the treatment effect. This is due to a loss of balance or randomization. The mean level of x_1 for the first strata (event number 1) is near 0 for both treatment and control. For strata 2, however, the mean levels were 0·5 and 0·1, respectively; high risk patients are more likely to have an event, and treated patients must be, on average, of higher risk than controls to have had one. This leads to the disparity between treatment groups with respect to x_1. By strata 5, the mean levels were 1·3 and 0·9, respectively. Use of 'time since last event' as the time scale changes the results only slightly.

The marginal analysis overestimates the treatment effect, and inclusion of the covariate x_1 into the model leads to serious overestimation. The problem here is that the data for strata 4, 5, etc. no longer obey the proportional hazards model. Stratum 3, for instance, contains all 500 subjects, so randomization over x_1 is not an issue. The time to first event, however, is the sum of three exponentials. It is easy to show that if X is exponential with hazard λ_1 and Y is exponential with rate λ_2, then the hazard ratio for the kth event is a decreasing function of time, from $(\lambda_1/\lambda_2)^k$ to an asymptotic value of λ_1/λ_2 as time goes to infinity. One should conduct per stratum checks of the proportional hazards assumption when utilizing the marginal approach. Figure 1 shows such a check for these data, utilizing the plot of Grambsch and Therneau.[20]

The difference between the marginal and conditional models is emphasized by a fit that contains the treatment by covariate interactions. In Table II, the variables rx_1, $rx_2, \ldots,$ rx_6+ refer to the estimated treatment effect within strata 1 to 5 and 6 or more, respectively. The stratum 1 estimate, for both models, is equal to the time to first event analysis. For further strata the marginal model's estimates rise and the conditional model's estimates fall. By stratum 6 the

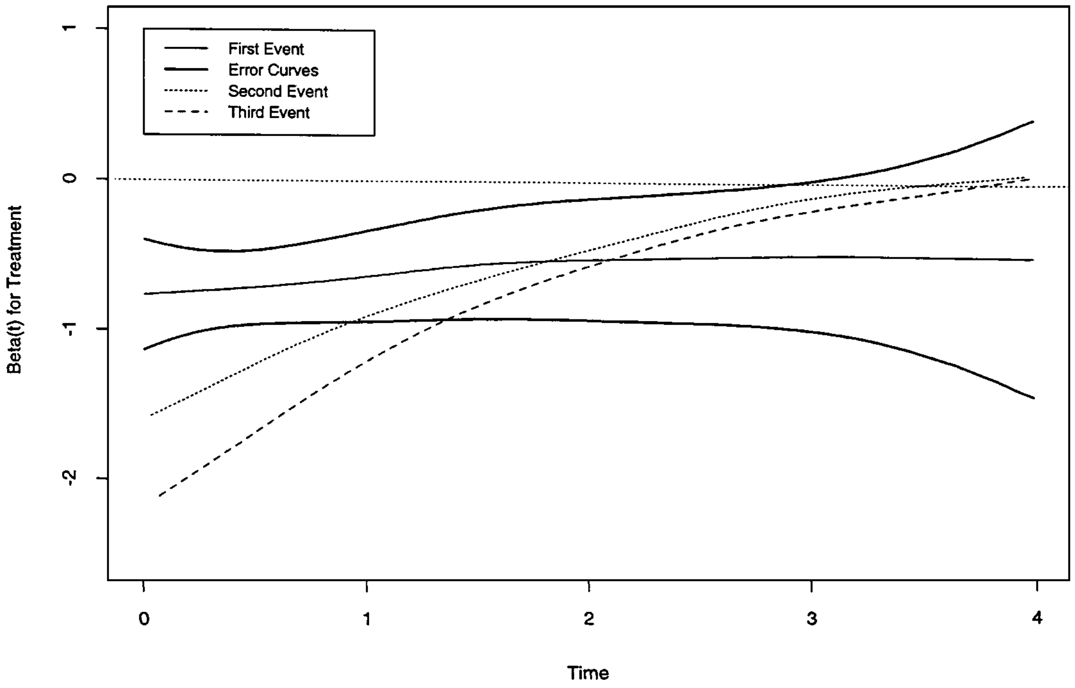

Figure 1. PH plots for simulated data, marginal model

Table II. Models with interaction

	rx_1	rx_2	rx_3	rx_4	rx_5	rx_6+
Marginal	− 0·89	− 1·12	− 1·38	− 1·71	− 2·00	− 2·63
Conditional	− 0·89	− 0·66	− 0·64	− 0·30	− 0·53	− 0·32

conditional estimates have become unstable; only 10 treatment subjects remain in the sample by that point.

Consider testing for an overall treatment effect. As discussed in Section 3.1, when the sequential events are not independent Cox models that depend on this assumption do not yield valid inference. Since, in this example, the sequential events were generated as independent the data satisfied this condition. If correctly specified, the AG model should be the most efficient, and the conditional model should be more efficient than the marginal model (Oakes[4]). When x_1 is known, the results support this, since the robust estimates of variance were smallest in the AG model and largest in the marginal. The naive estimates suggest the false conclusion that the conditional model is the least efficient. The efficiency of the AG model seemed the most sensitive to misspecification, since not including x_1 increased the variance of the AG model by 27 per cent. Lin[21] points out that the marginal model is the most robust for testing an overall treatment effect.

4.2. rIFN-g in Patients with Chronic Granulomatous Disease

Chronic granulomatous disease is a heterogeneous group of uncommon inherited disorders characterized by recurrent pyogenic infections that usually begin early in life and may lead to death in childhood. Interferon gamma is a principal macrophage-activating factor shown to correct partially the metabolic defect in phagocytes. It was hypothesized that treatment with interferon might reduce the frequency of serious infections in patients with CGD. In 1986, Genentech Inc. conducted a randomized double-blind placebo-controlled trial in 128 CGD patients who received Genentech's humanized interferon gamma (rIFN-g) or placebo three times daily for a year.[22] The primary endpoint of the study was the time to the first serious infection. However, data were collected on all serious infections until the end of follow-up, which occurred before day 400 for most patients. Thirty of the 65 patients in the placebo group and 14 of the 63 patients in the rIFN-g group had at least one serious infection. The total number of infections was 56 and 20 in the placebo and treatment groups, respectively. The full data set appears in appendix D·2 of Fleming and Harrington.[6] Lin[21] also compared multiple endpoint Cox models to analyse these data.

In choosing a model for the time to recurrent infections, the analyst should consider the biological processes of the disease. For instance, it is possible that after experiencing the first infection, the risk (*hazard*) of the next infection may increase. This could happen if each infection permanently compromised the ability of the immune system to respond to subsequent attacks. If this were the case, one would use a model containing separate strata for each event, or perhaps incorporate a time-dependent covariate. From practical experience, clinical scientists conducting the rIFN-g trial suggested that the risk of recurrent infection remained constant regardless of the number of previous infections. This suggests use of an independent increments or AG model.

Table III shows the results of several models. In the first model, time to first infection, the ordinary and robust variance estimate $D'D$ agree closely; a major disagreement would constitute evidence of violation of some assumptions of the Cox model. For the other fits, the robust variance column contains the grouped estimate $\tilde{D}'\tilde{D}$.

The Andersen–Gill model gives nearly an identical coefficient. If we ignore correlation between subjects, then there is an apparent reduction in variance of 39 per cent, from 0.112 to 0.068. Using the robust variance estimate $\tilde{D}'\tilde{D}$ the reduction is much smaller, only 13 per cent. This suggests that inclusion of all events in the analysis is worthwhile although the gain is slight.

The pattern of results for the marginal and conditional approaches is remarkably similar to the simulated example presented earlier. The conditional model results shown above are for time since entry; a fit using time since last event differs by only ± 0·01 from these. If we fit separate coefficients to the first three strata or event numbers, the results for the marginal model are $-1{\cdot}10$, $-1{\cdot}25$ and $-2{\cdot}74$, and for the conditional model they are $-1{\cdot}10$, 0·15 and $-1{\cdot}28$. (In the conditional setup, the treatment group has only five observations in stratum 3). Again, this is very similar to the hidden covariate example.

Because of this similarity, we might expect that the independent increment and conditional models would give closer results if the model included significant covariates. The two most important factors, other than treatment, are age and enrollment centre (the first three digits of the subject id). Table IV shows the results for the treatment effect in a model that includes both age and centre, the latter as a categorical variable with 13 levels. The results again parallel the hidden covariate data set; the AG and conditional coefficients have become closer in value. If we enter centre as a stratification variable, the results come even closer; the coefficients for the AG and conditional models are $-1{\cdot}23$ and $-1{\cdot}19$, respectively.

Table III. Fits for the CGD data

	β	SE(β)	robust SE
Time to first event	– 1·10	0·34	0·34
AG	– 1·10	0·26	0·31
Marginal	– 1·34	0·27	0·31
Conditional	– 0·87	0·28	0·29

Table IV. Fits for the CGD data, controlling for age and centre

	β	SE(β)	robust SE
Time to first event	– 1·25	0·35	0·35
AG	– 1·16	0·26	0·30
Marginal	– 1·51	0·28	0·37
Conditional	– 1·00	0·29	0·29

4.3. rhDNase in Patients With Cystic Fibrosis

In patients with cystic fibrosis, extracellular DNA is released by leukocytes that accumulate in the airways in response to chronic bacterial infection. This excess DNA thickens the mucus, which the cilia then cannot clear from the lung. The accumulation leads to exacerbations of respiratory symptoms and progressive deterioration of lung function. More than 90 per cent of cystic fibrosis patients eventually die of lung disease

Deoxyribonuclease I (DNase I) is a human enzyme normally present in the mucus of human lungs that digests extracellular DNA. Genentech Inc. has cloned a highly purified recombinant DNase I (rhDNase or Pulmozyme) which, when delivered to the lungs in an aerosolized form, cuts extracellular DNA, reducing the viscoelasticity of airway secretions and improving clearance. In 1992 the company conducted a randomized double-blind trial comparing rhDNase to placebo.[23] Patients were then monitored for pulmonary exacerbations, along with measures of lung volume and flow. The primary endpoint was the time until first pulmonary exacerbation; however, data on all exacerbations were collected for 169 days. The data is available on the StatLib.

Table V shows the results on the number of exacerbations. Overall 139/324 (43 per cent) of the placebo and 104/321 (32 per cent) of the rhDNase patients experienced an exacerbation during the follow-up period. A Cox proportional hazards model using the time to first exacerbation yields a hazard ratio of 0·69, with a 95 per cent confidence interval of (0·54, 0·89); strong evidence that rhDNase reduces the number of pulmonary events.

The data for second exacerbations, however, seem to point in the other direction: 42/139 (30 per cent) of the placebo and 39/104 (38 per cent) of the treated patients who had a first exacerbation went on to experience a second. We can use a multiple event Cox model to clarify and understand this result.

Since pulmonary exacerbations cause scar tissue to develop, which reduces lung function, it is reasonable to assume that the baseline hazard of each subsequent exacerbation was different. This suggests the use of either:

(i) a marginal model with one stratum per event number;

Table V. Distribution of exacerbations in rhDNase trial

Number of exacerbations	Placebo	rhDNase
0	185	217
1	97	65
2	24	30
3	13	6
4	4	3
5	1	0

(ii) a conditional model with one stratum per event number;
(iii) an Andersen–Gill model with a time dependent covariate for event number.

We should perhaps base the stratification on the total number of prior events rather than the number of prior events on this study, but this covariate is unavailable.

Setting up the data sets for these models was more complicated because of discontinuous intervals of risk. During an exacerbation, patients received intravenous (IV) antibiotics and were not considered at risk for a new event until seven exacerbation-free days beyond the end of IV therapy. Consider a single treated patient who had exacerbations at days 50 and 100 with antibiotic treatment durations of 10 and 15 days plus a seven day risk-free period following the end of IV antibiotic administration, and a final follow-up at day 180. We created two data sets to do the analysis. In the first, used for both the AG and conditional analysis, this patient would appear as three observations with data values:

time1	time2	status	rx	enum
0	50	1	1	1
67	100	1	1	2
122	180	0	1	3

For the marginal analysis, he will appear as 12 observations:

time1	time2	status	rx	enum
0	50	1	1	1
0	50	0	1	2
67	100	1	1	2
0	50	0	1	3
67	100	0	1	3
122	180	0	1	3
0	50	0	1	4
67	100	0	1	4
122	180	0	1	4
0	50	0	1	5
67	100	0	1	5
122	180	0	1	5

Table VI. Simple fits to the DNase data

	β	SE	robust SE	P
First event	− 0·365	0·13	0·13	0·005
Andersen–Gill	− 0·303	0·11	0·13	0·022
Marginal	− 0·345	0·11	0·15	0·019
Conditional	− 0·227	0·11	0·11	0·036

Table VII. Strata specific fits to the DNase data

	β	SE	robust SE	P
Marginal				
Baseline FEV_1	− 0·020	0·002	0·003	0·000
1st event	− 0·383	0·13	0·13	0·003
2nd event	− 0·092	0·22	0·22	0·680
> 2 events	− 0·737	0·35	0·43	0·089
Conditional				
Baseline FEV_1	− 0·015	0·002	0·003	0·000
1st event	− 0·380	0·13	0·13	0·003
2nd event	0·321	0·22	0·21	0·130
> 2 events	− 0·456	0·36	0·37	0·220

We determined that removing patients from the risk set for the 7 day risk-free interval had little effect on the estimates.

A further consideration is the very small number of events in strata 4 and 5. We have three possibilities to deal with this. The first is to treat them exactly like the other strata, accepting the fact that the within-strata hazard estimates are very unstable, perhaps even useless. This is particularly true for the conditional model, which has a very small sample size in this region. A second possibility is to truncate the data set after the third event. The third approach, which we have used, is to amalgamate. For the marginal model we may or may not preserve the strata; the important change is to model a single treatment effect for events 3–5. For the AG and conditional models, we affect the change by capping the strata variable enum at a value of 3.

Table VI shows the result of simple fits to the rhDNase data, and Table VII the result of more complicated models. The most acceptable model is the marginal model with separate treatment coefficients for each stratum, and a linear effect for the baseline level of forced expiratory volume in 1 second (FEV_1). In this we see an apparent lessening of the treatment effect in stratum 2 and an increase in stratum 3. The individual contrasts between rx_1/rx_2 and rx_1/rx_3 are not significant, however, with $p = 0{\cdot}18$ and 0·35, respectively. Given the problem with non-proportional hazards exhibited in the hidden variable example, it is best to test for this using the scaled Schoenfeld residuals. We observed no formal evidence for non-proportional hazards; however, a PH plot showed a worsening of the proportionality assumption in the higher strata.

We observed further effects of the loss of treatment group balance with the conditional model by fitting stratum specific coefficients for baseline FEV_1. In the first stratum, the baseline level of FEV_1 is a powerful predictor of an exacerbation. In the conditional model, exclusion of the

Table VIII. Estimated risk of exacerbation: double-blind and follow-up periods

Exacerbation number	Relative risk	Standard error
1	0·72	0·10
2	0·78	0·18
> 2	0·54	0·38

Table IX. Characteristics of patients; double-blind and follow-up periods

	Double-blind		Follow-up rhDNase
	Placebo	rhDNase	
Mean FVC (per cent)	78	78	78
Mean FEV_1(per cent)	61	61	61
Mean age	18	19	19

exacerbation-free patients from the risk set for a second event causes FEV_1 no longer to be predictive. In the marginal model, the effect of baseline FEV_1 is the same in each stratum.

The results of analysing the recurrent events from the double-blind trial suggested a possible diminishing treatment effect; but we could reach no conclusion as a result of too little information. However, the long term effect of rhDNase was estimable from data collected during a post-double-blind observational period. At the end of the 169 day trial, the treatment was determined efficacious and all participating patients were given rhDNase and followed for an additional 18 months. The cross-over of placebo patients to rhDNase was coded using a time-dependent treatment covariate, as in the Stanford Heart Study example. Standard errors are again based on the robust estimate $\tilde{D}'\tilde{D}$. Since the patients in the double-blind trial were enrolled during a six month period beginning in February 1992, data from months 7 to 12 of the follow-up period were used in the analysis to remove any seasonal effect. The data included observations from the placebo patients from the double-blind period and data from all patients during months 7 to 12 of the follow-up period. Table VIII summarizes the results of fitting a marginal model to the time until each exacerbation. The relative risk estimates of first and recurrent exacerbation suggest that the treatment effect during the follow-up was sustained and consistent with the double-blind period. Diagnostics did not indicate violation of the proportional hazards assumption.

Approximately 9 per cent of the patients dropped out during the extended follow-up period. If these patients were significantly more ill than the remaining patients, this could bias the previous comparison. To test this, we compared age and lung function at baseline to the values at month 7 of the follow-up; the results appear in Table IX. No difference in patient characteristics was seen.

5. SUMMARY

Multiple event studies, where the events are of the same type, are the most difficult for analysis. We believe that the most serious problem in assessing such studies is the loss of treatment balance when comparing later events. When coupled with the robust variance estimator $\tilde{D}'\tilde{D}$, the simple Andersen–Gill formulation appears the least affected by this problem, and should be our first choice for testing an overall treatment effect. If an evalution of the proportional hazards assumption does not show great departures, then one may use the WLW setup as well. It has the advantage of giving estimates of the possible change in treatment effect over time.

We can easily fit all three forms using current software, which should encourage further use and evaluation of the techniques.

APPENDIX I: DEFINITIONS

The Cox model assumes that the risk for subject i is

$$\lambda(t|Z_i) = \lambda_0(t) r_i(t)$$

where λ_0 is an unspecified baseline hazard. Assuming no tied death times, the log partial likelihood is defined as

$$l(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t) r_i(t) - \log\left\{ \sum_j Y_j(t) r_j(t) \right\} \right] \mathrm{d}N_i(t).$$

The first derivative is

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n} \int_0^\infty [Z_i(t - \bar{Z}(t)] \mathrm{d}N_i(t) \\ &= \sum_{i=1}^{n} \int_0^\infty [Z_i(t) - \bar{Z}(t)] \mathrm{d}M_i(t) \qquad (1) \end{aligned}$$

and the information matrix is a sum of weighted variance matrices

$$\mathscr{I} = \sum_{i=1}^{n} \int_0^\infty \frac{\sum_j Y_j(t) r_j(t) [Z_j(t) - \bar{Z}(t)][Z_j(t) - \bar{Z}(t)]'}{\sum_j Y_j(t) r_j(t)} \mathrm{d}N_i(t) \qquad (2)$$

where $\bar{Z}$ is the weighed mean of those still at risk at time t

$$\bar{Z}(t) = \frac{\sum Y_j(t) r_j(t) Z_j(t)}{\sum Y_j(t) r_j(t)}.$$

The Breslow estimate of the baseline hazard (also refered to as the Link, Tsiatis, or Nelson–Aalen estimate) is

$$\hat{\Lambda}_0(\beta, t) = \sum_{i=1}^{n} \int_0^t \frac{\mathrm{d}N_i(s)}{\sum_{i=1}^{n} Y_i(s) r_i(s)}.$$

APPENDIX II: SAS AND S-PLUS CODE

The following shows the fit and output for models fit to the CGD data. There are two data sets: `gamma` is set up in the AG style and has 203 observations on 128 patients, `gamma2` contains the

WLW style data and has $128 \times 7 = 896$ observations. The counting process style of computation shown below requires version 3.3 or later of S-plus, and version 6.10 or later of SAS.

S-plus

The ' > ' character at the beginning of a line represents the S-plus prompt. Text following the prompt is typed by the user. The fits are afit = Andersen–Gill model, mfit = marginal model, cfit = conditional model, and cfit2 = a conditional model that includes the treatment by strata interaction. The printout has been omitted for all but the first model.

```
  > afit ← coxph(Surv(t1, t2, status) ~ rx + age + cluster(id),
  data = gamma)
  > summary(afit)
    n = 203
    robust variance based on 128 groups

          coef  exp(coef)  se(coef)  robust se       z        p
rx   - 1.1201       0.326    0.2613     0.3099   - 3.61   0.0003
age  - 0.0305       0.970    0.0131     0.0144   - 2.11   0.0340

     exp(coef)  exp( - coef)  lower .95  upper .95
rx       0.326          3.07      0.178      0.599
age      0.970          1.03      0.943      0.998

Rsquare = 0.12   (max possible = 0.966)
Likelihood ratio test = 25.9 on 2 df,   p = 2.38e - 06
Wald test             = 16.6 on 2 df,   p = 0.000246
Efficient score test  = 24.8 on 2 df,   p = 4.05e - 06

  (Note: the likelihood ratio and efficient score tests assume independence of the
observations).

 > cfit ← coxph(Surv(t1, t2, status) ~ rx + strata(enum) + cluster(id), data = gamma)

 > mfit ← coxph(Surv(time, status) ~ rx + strata(enum) + cluster(id), data = gamma2)

 > cfit2 ← coxph(Surv(t1, t2, status) ~ rx * strata(enum) + cluster(id), data = gamma)
```

SAS

Although creation of the gamma and gamma2 data sets themselves is easier to do in SAS, the analysis is a bit more clumsy. Below we only show the first (agfit) of the S-plus models above. First we fit the model ignoring the correlation. Then we form $\tilde{D}$ from D using proc means, then we form the product $\tilde{D}'\tilde{D}$ using proc iml, and finally we print the robust variance. We could gather together all of these steps using a macro, of course, to give a unified call and printout similar to the S-plus function, but allowing for all of the phreg options in such a macro is tedious and we do not pursue it here. We must add interaction variables to the data set, if desired, prior to the use of phreg.

```
proc phreg data = gamma;
     model (t1, t2) * status(0) = rx age;
     output out = temp1 ressco = res_rx res_age;
     id id;

proc sort data = temp1; by id;
proc means;
     by id;
     var res_rx res_age;
     output = temp2 sum = rx age;

proc iml;
     use temp2; setin temp2;
     read all var rx age into r;

     rvar = r' * r;
     create temp3 from rvar[rowname = rx age   colname = rx age];
     quit;

proc print data = temp3;
```

REFERENCES

1. 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).
2. 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).
3. Prentice, R.L., Williams, B.J. and Peterson, A.V. 'On the regression analysis of multivariate failure time data', *Biometrika*, **68**, 373–389 (1981).
4. Oakes, D.A. 'Frailty models for multiple event times', *in* Klein, J.P. and Goel, P.K. (eds), *Survival Analysis*, *State of the Art*, Kluwer Academic Publishers, Netherlands, 1992, pp. 415.
5. Prentice, R.L. and Cai, J. 'Covariance and survivor function estimation using censored multivariate failure time data', *Biometrika*, **79**, 495–512 (1992).
6. Fleming, T.R. and Harringon, D.P. *Counting Processes and Survival Analysis*, Wiley, New York, 1991.
7. Therneau, T.M., Grambsch P.M. and Fleming, T.R. 'Martingale based residuals for survival models', *Biometrika*, **77**, 147–160 (1990).
8. Barlow, W.E. and Prentice, R.L. 'Residuals for relative risk regression', *Biometrika*, **75**, 65–74 (1988).
9. Cain, K.C. and Lange, N.T. 'Approximate case influence for the proportional hazards regression model with censored data', *Biometrics*, **40**, 493–499 (1984).
10. Reid, N. and Crépeau, H. 'Influence functions for proportional hazards regression', *Biometrika*, **72**, 1–9 (1985).
11. Crowley, J. and Hu, M. 'Covariance analysis of heart transplant data', *Journal of the American Statistical Association*, **72**, 27–36 (1977).
12. White, H. 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica*, **48**, 817–830 (1980).
13. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrika*, **50**, 1–25 (1982).
14. Liang, K.L. and Zeger, S.L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
15. Lin, D.Y. and Wei, L.J. 'The robust inference for the Cox proportional hazards model', *Journal of the American Statistical Association*, **84**, 1074–1079 (1989).
16. Lee, E.W., Wei, L.J. and Amato D. 'Cox-type regression analysis for large number 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, Netherlands, 1992, pp. 237–247.
17. Lunn, A.D. and McNeil, D.R. *Computer-interactive Data Analysis*, Wiley, New York, 1991.

18. Laird, N.M. and Olivier, D. 'Covariance analysis of censored survival data using log-linear analysis techniques', *Journal of the American Statistical Association*, **76**, 231–240 (1981).
19. Omori, Y. and Johnson, R.A. 'The influence of random effects on the unconditional hazard rate and survival function', *Biometrika*, **80**, 910–924 (1993).
20. Grambsch, P. and Therneau, T.M. 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994).
21. Lin, D.Y. 'Cox regression analysis of multivariate failure time data', *Statistics in Medicine*, **15**, 2233–2247 (1994).
22. The International Chronic Granulomatous Disease Cooperative Study Group. 'A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease', *New England Journal of Medicine*, **324**, 509–516 (1991).
23. Fuchs, H.J., Borowitz, D., Christiansen, D., Morris, E., Nash, M., Ramsey, B., Rosenstein, B.J., Smith, A.L. and Wohl, M.E. 'The effect of aerosilozed recombinant human DNase on respiratory exacerbations and pulmonary function in patients with cystic fibrosis', *New England Journal of Medicine*, **331**, 637–642 (1994).