Extending the Cox Model

Terry M. Therneau

Technical Report Number 58
November 1996

Technical Report Series

Section of Biostatistics

Mayo Clinic, Rochester, Minnesota

Contents

1 Introduction

Since its introduction, the proportional hazards model proposed by Cox [8] has become the workhorse of regression analysis for censored data. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory. Comprehensive accounts of the underlying mathematics are given in the books of Fleming and Harrington [14] and of Andersen et. al. [2]. These developments have, in turn, led to the introduction of several new extensions of the original model. These include the analysis of residuals, time varying covariates, time dependent coefficients, multiple/correlated observations, multiple time scales, time dependent strata, and estimation of underlying hazard functions.

The aim of this monograph is to show how many of these methods and extensions of the model can be approached using standard statistical software, in particular the S-Plus and SAS packages. As such, it should be a bridge between the statistical journals and actual practice. The focus on SAS and S-Plus is based largely on the author's familiarity with these two packages, and should not be taken as evidence against the abilities of other software products. The text uses the labels 'S' and 'S-Plus' interchangeably; the former is the package developed by Bell Laboratories and the latter is the commercial version of the same. Since nearly every installation of "S" consists of the latter this shorthand notation should cause no harm. All of the examples given in the text actually refer to S-Plus.

Sections 2 and 3 lay a foundation for our methods. In section 2, we discuss the *counting process* formulation of a Cox model, the software implementation of this model, and the flexibility that it allows. Section 3 defines a set of residuals for the Cox model, based on the counting process and the mathematical formalism of a martingale.

Sections 4 and 5 use residuals to test the two basic assumptions of a Cox model: proportional hazards and the functional form of the covariates. The counting process formulation allows us to extend these methods to time-dependent covariate models as well.

Section 6 discusses one of the newer areas of application, the use of a Cox model for correlated survival data. Such data naturally arise when there are multiple observations per subject as well as in other applications. Because of it's importance, and several choices that are available in the set up of such problems, more examples are presented in this area than in any of the other sections.

2 The counting process formulation of a Cox model

The Andersen-Gill (AG) formulation of the proportional hazards model as a counting process has proven very useful in theoretical development [1]. Represent the ith subject as a *counting process* where

- $N_i(t)$ is the cumulative number of events up to time t for the subject.
- $Y_i(t)$ is an indicator function, $Y_i(t) = 1$ if and only if the subject is at risk of an event and under observation at time t.

From a data analysis viewpoint, each subject is treated as an observation of a (very slow) Poisson process. A censored subject is thought of not as "incomplete data", but as one whose event count is still zero. Time dependent covariates effect the rate for upcoming events, and can depend in any way on past observation of the subject. Intervals of observation need not be contiguous.

To cast a data analysis in this framework has several advantages. In the computer data set, each subject i is represented by a set of observations: s_{ij}, t_{ij}, δ_{ij}, x_{ij}, k_{ij}, $j = 1, \ldots, n_i$; where $(s_{ij}, t_{ij}]$ is an interval of risk, open on the left and closed on the right, $\delta_{ij} = 1$ if the subject had an event at time t_{ij}, and 0 if the subject did not have an event, x_{ij} is the covariate vector over the interval, and k_{ij} is the stratum to which the subject belongs during the interval. Data sets like this are easy to construct with a package such as SAS or S-Plus.

2.1 Particular cases

2.1.1 Multiple events

The original motivation for adding the counting process form to the S-Plus `coxph` function was a study of a calcium channel blocker, diltiazem, in post myocardial infarction patients, where one of the events of interest was fatal or non-fatal re-infarction [37]. Several patients had multiple events, though none had more than 3. A "standard" Cox analysis was performed using first cardiac event as an endpoint, but there was a question about whether more power would be obtained if all of the events could be used. These additional data can be incorporated by breaking any patient with multiple events into multiple intervals of risk. For instance, 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], (185, 250] with corresponding status codes of 1, 1 and 0.

2.1.2 Time-dependent covariates

The most common type of time dependent covariate is a repeated measurement 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 enrollment to death are 102 and 343 days, respectively; the second patient had a transplant 21 days from enrollment. The data file for these two patients would be

Interval	Status	Transplant	Age at Entry	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.

Multiple lab values are easily coded as well. A patient with tests on days 0, 60 and 120, and follow up to day 140 would be coded using 3 time intervals 0–60, 60–120 and 120–140. This does implicitly assume that the time dependent covariate is a step function with jumps at the measurement points. It might be more reasonable to break at the midpoints of the measurement times, or to use an interpolated value over many smaller intervals of time, but in practice these refinements appear to make little practical difference in a model's results. If a lab test varies *markedly* from visit to visit interpolation strategies may become important, but the adequacy of the study design would then also be in question.

2.1.3 Discontinuous intervals of risk

In a study of tumor progression and its relationship to a particular blood marker, the key time-dependent variable was the monthly measurement of this marker. Patients were expected to have a measurement every 3-6 months. One patient, however, had a significant hiatus in her visit record. One choice for the analysis was to interpolate the values over the missing 2 year period. A more conservative course, and that chosen by the investigators, was to treat the data value as missing. This effectively removes the subject from the risk set over that interval, without having to remove her entire experience from the study.

Other cases where a subject is lost from and then returns to observation can certainly be imagined. For instance, consider a situation where multiple events are possible, but the treatment for an event temporarily protects the patient from further injury. In a study of falls in the elderly, hospitalization following a fall would temporarily protect the subject from further falls. (For conditions with a low event rate, however, this refinement is likely to be insignificant.)

2.1.4 Alternate 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.

Andersen et al. [2] discuss this issue in several of their examples, and then in depth in chapter 10 of their book. One example concerns nephropathy and mortality among insulin dependent diabetics. Patients can be in one of 3 states: 0-alive without diabetic nephropathy (DN), 1-alive with DN, and 3-dead. Relevant time scales for the 0-1 transition are age, calendar time, and duration of diabetes, and for the 1-2 transition the duration of DN.

As another example, consider a study conducted at the Mayo Clinic on the effect of L-dopa for patients with Parkinson's disease. It was felt that time from diagnosis was the most appropriate time scale for analysis. However, Mayo is a major tertiary referral center, and many of the study's patients were not seen here until well after that date. For each patient we have the date of diagnosis, the date of referral to Mayo, and the date of last follow-up or death. The patient diagnosed on 8Feb82 with referral on 28Apr85 and last contact on 18Jun90 will be represented as a single interval (1175, 3052]. It is not correct to enter them as (0, 3052] (which is equivalent to a standard non-interval Cox model with follow-up time of 3052) since the patient was not at risk for an *observable* death during that interval. Such data, where the patient enters the risk set after time 0, is said to be left truncated.

2.1.5 Time dependent strata

When a patient is represented as multiple lines of data or "observations", both the covariates and the stratum indicator may change from line to line. Coding a time dependent stratum is thus quite easy.

Time alignment within the strata may require more thought, however. As an example, consider a study of Dutch patients with primary biliary cirrhosis of the liver (PBC). PBC is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50 cases per million population. The hazard rate for a diseased patient grows over time, as does the rate of degeneration in their hepatic function as tracked by various blood tests. A portion of the patients receive a liver transplant at some point during their follow up.

One point of the study was to assess the value of covariates such as age and bilirubin in predicting patient outcome, both before and after transplantation. Transplant was treated as a time dependent stratification variable. In the post transplant stratum the most "natural" hazard function is based on time since transplant. Surgical death is a major risk for such an extensive procedure, and this time scale properly aligns the patient's clock with the dominating hazard.

The "proper" alignment for time dependent strata is not always so clear. One appealing method of analysis for the diltiazem study is to place patients into new strata after their second, third, etc cardiac event (all have had one event, which was the trigger for enrollment). The baseline hazard after a second infarction may be quite different than the group as a whole. It is not obvious, however, whether time since enrollment or time since last event is the better index of an appropriate risk group.

2.2 Implementation in SAS and S-Plus

The S-Plus function `coxph` and the SAS `phreg` procedure accommodate these extensions by a simple programming artifice. The input data set is assumed to consist of observations or rows of data, each of which contains the covariate values Z, a status indicator 1=event 0=censored and an optional stratum indicator, along with the time interval $(start, stop]$ over which this information applies. In the notation above, this means that each row is treated as a separate subject whose Y_i variable is 1 on the interval $(start, stop]$ and zero otherwise. Within the program, it means that the risk set at time t only uses the applicable rows of the data. In order to avoid double-counting any subject within a risk set, both packages disallow zero length intervals, i.e. $(x, x]$. At the time of this writing S-Plus will fail with an error message if such intervals are present, and SAS silently deletes them. I slightly prefer the former behavior since the presence of such intervals is often indicative of a more general mistake in setting up the data set.

The code has no specific "hooks" to accommodate time-dependent covariates, time-dependent strata, multiple events, or any of the other special features mentioned above. Rather, it is the responsibility of the user to first construct an appropriate data set. The strategy, originally motivated by sloth, leads to a fitting program that is simpler, shorter, easier to debug, and more efficient than one with multiple specific options. A significantly more important benefit has become apparent over time, i.e., the flexibility inherent in building a data set has allowed analyses that were not considered by the original coder — left truncation is a case in point.

The S-Plus code has had the ability to fit these models since version 2.0 (1992) using the `agreg` function; from S-Plus version 3.3 (1995) onward it is easier to use the `coxph` function. The counting process form is an option within `phreg` for SAS releases 6.10 and later; the actual date of availability depends on your specific computing platform.

2.2.1 The Stanford Heart Transplant Data

Since the Stanford data set is so well known, it is worthwhile to show the data setup in complete detail. We use the data set as it is found in the paper by Crowley and Hu [11]. A copy of the data can be obtained from statlib (http://lib.stat.cmu.edu) in the *jasa* section.in the *jasa* section. Three issues arise in setting up the data:

1. The covariates in the data set are moderately colinear. Because of the presence of an interaction term, the coefficients found in table 5.2 of Kalbfleisch and Prentice [22] will be recreated only if the covariates are defined in *exactly* the correct way. (Conclusions are not changed, however). This is the reason for using fractional age (days/365.25), centered at 40 years, and for centering the enrollment date at Oct 1, 1967.

2. One subject died on the day of entry. However $(0, 0]$ is an illegal time interval for the programs. To avoid this, treat an interval from 10/21 to 10/22, say, as $\underline{2}$ days of risk, i.e., someone enrolled on 10/21 and censored on 10/22 should be coded as the interval $(0, 2]$. This leads to the "`1 +`" expression in several lines of the program.

3. A subject transplanted on day 10 is considered to be on medical treatment for days 1–10 and on surgical treatment for days 11–last contact. Thus if Smith died on day 10 and Jones was transplanted on day 10, we in effect treat the transplant as happening later in the day than the death; *except* for patient 38, who died during surgery on day 5. This person should certainly be counted as a treatment death rather than a medical one. The problem is resolved by moving his transplant to day 4.9. In the final data set his first observation is $(0, 5.9]$ with status of 0 and a treatment of 'non-surgical' over the interval (remember the +1 day rule), and a second observation of $(5.9, 6]$ with a status of 1 (dead) and surgical treatment.

(If there are ties between the time at which some time dependent covariate changes value and an event or censoring time, I have usually found it most accurate to explicitly resolve the ambiguity via a fractional time value, rather than try to remember and apply any tie breaking rule used by the program code. However, for those who wish to know there are two principles. Since time intervals are open on the left and closed on the right, changes in a covariate by default happen *after* the deaths and/or censoring at a time point. For ties between death and censoring times SAS and S-Plus place deaths first, in accordance with common practice.)

Here is SAS code to create the analysis data set.

```
data temp;
```

```
    infile 'data.jasa';
    input id  @6 birth_dt mmddyy8. @16 entry_dt mmddyy8.
              @26 tx_dt mmddyy8.   @37 fu_dt mmddyy8.
              fustat  prior_sx ;

    format birth_dt entry_dt tx_dt fu_dt date7.;

data stanford;
    set temp;
    drop fu_dt fustat birth_dt entry_dt tx_dt;

    age =  (entry_dt - birth_dt)/365.25  - 48;
    year = (entry_dt - mdy(10,1,67))/ 365.25;   *time since 10/1/67;
    wait =  1 + (tx_dt - entry_dt);
    if (id = 38) then wait = wait - .1;

    if (tx_dt =.) then do;
        rx = 0;               * standard therapy;
        start = 0;
        stop  = 1 + fu_dt - entry_dt;
        status= fustat;
        output;
        end;

    else do;
        rx =0;         *first an interval on standard treatment;
        start = 0;
        stop  = wait;
        status= 0;
        output;

        rx =1;         *then an interval on surgical treatment;
        start = wait;
        stop  = 1 + fu_dt - entry_dt;
        status= fustat;
        output;
        end;

proc print;
    id id;
```

And here is S-Plus code to fit the 6 models found in table 5.2 of Kalbfleisch and Prentice [22]. The > symbol is the package's interactive prompt. The formula language uses $\sim$ for "is modeled as", and the right hand side symbols are similar to the GLIM and GENSTAT programs: + for main effects, : for interaction, and

* for main effects plus interaction. The contrast option that I have chosen causes it to use the first treatment as the reference category.

```
> options(contrasts="contr.treatment")

> sfit.1 <- coxph(Surv(start, stop, status) ~ (age + prior.sx)* rx,
                               data=stanford, method='breslow')
> print(sfit.1)
               coef exp(coef) se(coef)      z    p
        age  0.0139     1.014   0.0181  0.768 0.44
   prior.sx -0.5465     0.579   0.6109 -0.895 0.37
         rx  0.1195     1.127   0.3277  0.365 0.72
     age:rx  0.0346     1.035   0.0272  1.270 0.20
prior.sx:rx -0.2929     0.746   0.7582 -0.386 0.70

Likelihood ratio test=12.5  on 5 df, p=0.0288  n= 172

> sfit.2 <- coxph(Surv(start, stop, status)~ year* rx,
                               data=stanford, method='breslow')

> sfit.3 <- coxph(Surv(start, stop, status)~ (age + year)* rx,
                               data=stanford, method='breslow')

> sfit.4 <- coxph(Surv(start, stop, status)~ (year +prior.sx) *rx,
                               data=stanford, method='breslow')

> sfit.5 <- coxph(Surv(start, stop, status)~ (age + prior.sx)* rx +
                          year, data=stanford, method='breslow')

> sfit.6 <- coxph(Surv(start, stop, status)~ age* rx + prior.sx +year,
                               data=stanford, method='breslow')
```

Because of several tied death times in the data set, the Efron approximation would be preferred to the Breslow method, but the difference in this case is slight and we wished to match the printed table. (The Efron approximation is the default in S-Plus). The SAS code to analyse the data set requires explicit creation of the interaction variables. An example for the first of the models is given below

```
data temp2; set stanford;
    age_rx   = age * rx;
    prior_rx = prior_sx * rx;
proc phreg data=temp2;
    model start stop * status(0) = age prior_sx rx age_rx prior_rx;
```

2.3 Computational considerations

The more common way to deal with time dependent covariates is to call a computational subroutine at each death time. The counting process style has been implemented in the S-Plus package, and more recently in SAS as well. There are some advantages for the procedure proposed here.

1. It was easy to code the routine.

2. The counting process code may be more efficient. Both routines begin by sorting the data according to the end point of the interval. In both, the time needed to accumulate the score and information matrices is identical. The difference is that the counting process routine must spend time scanning the data to select the risk set — those observations whose (start, stop] interval brackets a given death time. The other style routine must make n_l subroutine calls, where n_l is the size of the risk set. If the subroutine is at all complex the tradeoff will favor the counting process routine, particularly if only a few subjects have multiple lines of data. For example, in a recent study here with multiple lab tests, the proportion with 1, 2, . . . tests geometrically decreased; only 2 patients had five values. Slightly over half the subjects had only one value.
   The counting process function will often run much faster when there are stratification variables in the model. When strata are introduced the program spends less time searching out who is part of the current risk set since it need look only within the strata; without strata it has to scan the entire data set.

3. The counting process style may be easier to use. This is particularly true for the case of multiple lab tests. Assume that there is a maximum of 5 lab tests, then use of the older style code requires the creation of 9 ancillary covariates containing the first lab value, the time of the second test, the value of the second test, the time of the third test, the value of the third test, etc. Those with fewer than 10 tests have the remaining "time of test" values set to some number greater than that patient's follow up time. Then a code fragment similar to the following is placed into the subroutine:

   ```
   lab = lab1;
   if (TIME > t1) then lab=lab2;
   if (TIME > t2) then lab=lab3;
   ...
   ```

   As well, if some patient has the effrontery to come for an 11th visit, then the subroutine and *all* ancillary data must be updated to include a new pair of variables.

The major disadvantage of the counting process style is that it is difficult to accommodate a smoothly time-varying covariate. One main use of this is to test the proportional hazards assumption by the addition of a product variable `x*time` or `x*log(time)`. As we will see later, there is a much better way to test for and visualize non-proportional hazards.

This is not to say that one *cannot* accommodate a smooth covariate. With sufficient effort, general time dependent covariates can be managed in any package that allows stratified models. This may be accomplished by making each unique death time a separate stratum, then within the stratum place one observation for each subject at risk, setting his/her covariates to the appropriate time-dependent values. Creation of such a data set would of course be a tedious process.

3 Residuals

3.1 Mathematical definitions

Several ideas for residuals based on a Cox model have been proposed on an ad hoc basis, most with only limited success. The current, and most successful, methods are all based on counting process arguments, and in particular on the subject-specific martingale process that arises from this formulation. Barlow and Prentice [3] give the original definition of "martingale residuals", and further work was done by Therneau, Grambsch and Fleming [47]. To begin, we must expand on the mathematical notation introduced in the previous section.

As before, let $Y_i(t)$ be the indicator that a given subject is at risk and under observation at time t, $N_i(t)$ the cumulative number of "events" for the subject up to time t, and $\overline{N}(t) = \sum N_i$ the total event process. Let $Z_{ij}(t)$ be the jth covariate of the ith person (possibly time dependent), where $i = 1, \ldots, n$ and $j = 1, \ldots, p$; and $Z_i(t)$ the entire covariate set for a subject i, represented as a $p \times 1$ column vector. Define $r_i(t)$ to be $\exp[\beta' Z_i(t)]$, i.e., the risk score for the ith subject. In actual practice β will be replaced by $\hat{\beta}$ and the subject weights r_i by $\hat{r}_i$.

If $\Lambda_i(t)$ is the true cumulative event rate function for subject i, then the (observed - expected) process

$$M_i(t) = \int Y_i(t)\{dN_i(t) - d\Lambda_i(t)\}$$

will be a subject-specific martingale.

The Cox model assumes that the hazard 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

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

The first derivative is the p by 1 vector

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n} \int_0^\infty [Z_i(t) - \bar{Z}(\beta,t)]\, dN_i(t) && (1) \\ &= \sum_{i=1}^{n} \int_0^\infty [Z_i(t) - \bar{Z}(\beta,t)]\, dM_i(\beta,t)\,. && (2) \end{aligned}$$

M is defined below. Equation 1 is the usual form in which the Cox score is written, but 2 is more useful in some contexts. The transition from 1 to 2 involves simple algebraic manipulations The p by p information matrix is

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \int_0^\infty \frac{\sum_j Y_j(t) r_j(t) [Z_i(t) - \bar{Z}(t)][Z_i(t) - \bar{Z}(\beta,t)]'}{\sum_j Y_j(t) r_j(t)} dN_i(t)\,, \qquad (3)$$

where $\bar{Z}$ is the weighted mean (vector) of those still at risk at time t: defined by

$$\bar{Z}(\beta,t) = \frac{\sum Y_i(t) r_i(t) Z_i(t)}{\sum Y_i(t) r_i(t)}\,. \qquad (4)$$

In practice the martingale residual is estimated by $\widehat{M}$, with $\hat{\beta}$, $\hat{r}$ and $\hat{\Lambda}$ substituted for β, r, and Λ. The most usual estimate of the baseline hazard is the Breslow or Nelson–Aalen estimate

$$\hat{\Lambda}_0(t) = \int_0^t \frac{d\overline{N}(s)}{\sum_{j=1}^{n} Y_j(s) \hat{r}_j(s)}.$$

3.2 Martingale and deviance residuals

The martingale residual for subject i at time t is

$$M_i(t) = N_i(t) - \int_0^t r_i(\beta,s) Y_i(s) d\hat{\Lambda}_0(\beta,s).$$

The SAS and S-Plus functions return the residual at $t = \infty, \beta = \hat{\beta}$. If there are no time-dependent covariates, then $r_i(t) = r_i$ and can be factored out of the integral, giving $\widehat{M}_i \equiv \widehat{M}_i(\infty) = N_i - \hat{r}_i \hat{\Lambda}_0(\hat{\beta}, t_i)$.

The martingale residual is really a simple difference, $O-E$, between the observed number of events for a subject and the expected number given the current model. As in a simple Poisson model, $\text{var}(M_i) = E_i$, the expected number of events. The residuals have some properties that are reminiscent of the residuals from an ordinary linear model.

1. $E(M_i) = 0$
2. $\sum \widehat{M}_i = 0$
3. $\text{cov}(M_i, M_j) = 0$
4. $\text{cov}(\widehat{M}_i, \widehat{M}_j) < 0$.

The last property is a consequence of the fact that the observed residuals are constrained to sum to zero. The covariance is small, however; of the same order as the covariance of residuals from a linear model.

The martingale residual is highly skewed for single-event (survival) data, since the limiting distribution (under the true model) of $\Lambda_0(t)$ is a censored exponential distribution. This suggests the application of a normalizing transform for the residual, with the idea that such plots would be "easier on the eye". The deviance residual incorporates such a normalizing transform and is very similar in form to the deviance residual for a Poisson distribution,

$$d_i = \text{sign}(\widehat{M}_i) * \sqrt{-\widehat{M}_i - N_i \log((N_i - \widehat{M}_i)/N_i)}$$

In practice it has not been very useful.

3.3 Score residuals

The score process $U_{ij}(t)$ for the ith subject and the jth variable is

$$U_{ij}(t) = \int_0^t [Z_{ij}(s) - \bar{Z}_j(s)]\, d\widehat{M}_i(s),$$

where $\bar{Z}_j(s)$ is the weighted mean of the jth covariate over the subjects still in the risk set at time s, as defined in equation 4.

One can think of the score process as a 3 way array with dimensions of subject, covariate, and time. Lin, Wei and Ying [32] suggest a global test of the proportional hazards model based on the maximum of the array.

The *score residual* is defined as a $U_{ij}(\infty)$, and is an n by p matrix. It is the sum of the score process over time. The first, and most obvious use of the score residual is as a measure of leverage. The *dfbeta* matrix is defined as $D = -U\mathcal{I}^{-1}$, where

$\mathcal{I}^{-1}$ is the usual Cox variance matrix. The elements of D are the leverage residuals derived by Cain and Lange [5] using case weights, and by Reid and Crépeau [44] using another method. The ij element D_{ij} is an estimate of the change in $\hat{\beta}_j$ if observation i were removed 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. A useful variant is *dfbetas*, the change in the coefficients scaled by the standard errors of the coefficients. It is common to plot either D_i (dfbeta) or dfbetas against $i = 1, 2, \ldots, n$ as a visual way of checking influence.

A second use of the leverage residuals matrix is to form a robust variance estimate. This will be discussed further in the section on multiple events.

3.4 Schoenfeld residuals

Another useful transform is to sum the score process up over individuals to get a total score process $L_j(t)$ which is a function of time. This yields a residual first proposed by Schoenfeld [45]. Because $\hat{\Lambda}_0$ is discrete, our estimated score process will also be discrete, having jumps at each of the unique event times. The Schoenfeld residuals are a p column matrix with one row per death which contains these increments.

3.5 General residuals

Barlow and Prentice [3] define a generalized martingale residual by

$$e_i(f_i) = \int f_i(t)\, d\widehat{M}_i(t)$$

for a predictable process f_i, which is any function of time or the covariates that depends only on the past. They show that since the predictable covariation process for M_i and M_j is zero, $e_i(f_i)$ and $e_j(f_j)$ are asymptotically uncorrelated. The variance process is

$$\mathrm{var}(e_i(f_i)) = \int f_i(t)[f_i(t)]'\, d\Lambda_i(t)$$

This formula can be used to motivate several residuals. The martingale residual is $e(1)$ and the score residual is $e(Z - \bar{Z})$. Other possibilities suggested by the authors are $e(t)$, "a more traditional residual giving the difference between the failure time t_i for an uncensored subject and a corresponding projected quantity under the model, and $e(\eta)$ where η is the linear predictor $Z'\beta$.

3.6 Counting process data

Assume that we have a counting process style of data set, where each subject may be represented by multiple (start, stop] intervals. One nice feature of the martingale and score residuals is that the residual for a subject is the sum of the residuals for his/her observations. This is a natural consequence of the integral representation. For example, assume that subject "Smith" is represented in the data as four observations with intervals (0,10], (10,25], (25,27] and (27,50]. The martingale residuals for the individual observations are

$$\int_0^{10} Y_i(s)\, d\widehat{M}_i(s)\,, \quad \int_{10}^{25} Y_i(s) d\widehat{M}_i(s)\,,$$
$$\int_{25}^{27} Y_i(s)\, d\widehat{M}_i(s)\,, \text{ and } \int_{27}^{50} Y_i(s)\, d\widehat{M}_i(s)\,.$$

The martingale residual for the subject is the integral from 0 to 50, which is obviously the sum of these four quantities.

The Schoenfeld residuals are defined at each unique death, and are unchanged by a counting process formulation.

4 Functional form

In the Cox model, we assume that the hazard function satisfies

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' Z).$$

That is, a proportional hazards structure with a log-linear model for the covariates. In this section we investigate the correct functional form for the covariates. Perhaps one of the variables, Z_j, should be replaced by Z_j^2, $\ln Z_j$, $I_{\{Z_j > c\}}$ or some other transform to properly account for its effect.

We use the PBC data set as an example. The data come from a Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984. PBC is a progressive disease thought to be of an auto-immune origin; the subsequent inflammatory process eventually leads to cirrhosis and destruction of the liver's bile ducts. A description of the study along with a listing of the data set can be found in Fleming and Harrington [14]. A more extended discussion can be found in Dickson, et al. [12]. Through the work done for these two analyses, the important variables and their proper transformation is already 'known'.

4.1 A Simple approach

The simplest approach is one examined by Therneau, Grambsch and Fleming [47] who suggested smoothed residual plots. Consider the martingale residuals from a null model, i.e., one with $\beta = 0$. They show that if the correct model is $\exp(\beta f(Z))$

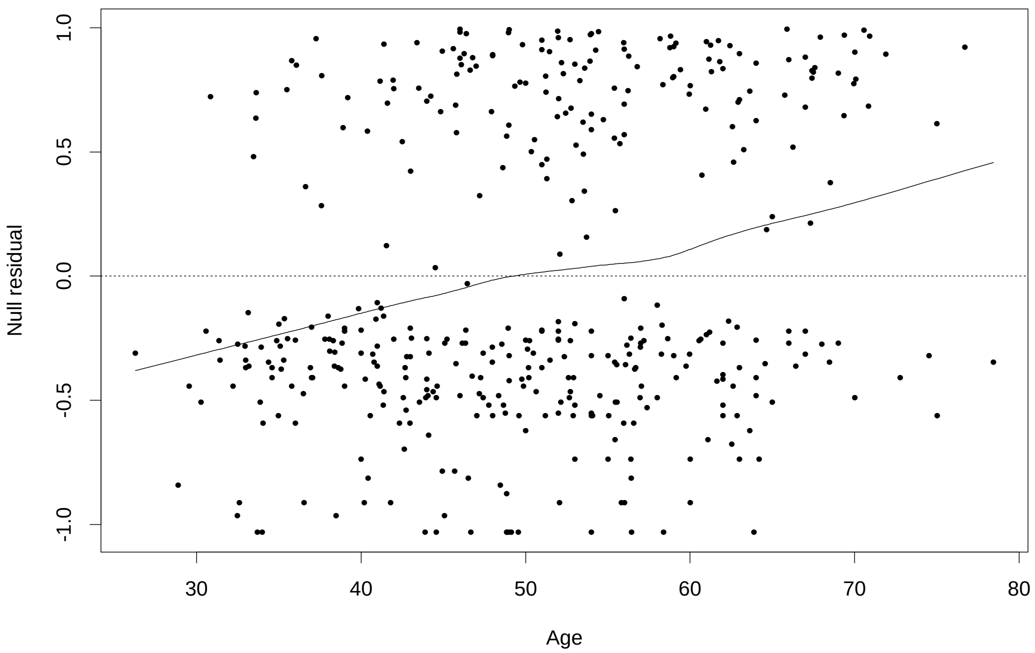

Figure 1: PBC Data, functional form for age

for some smooth function f, then a plot of the smoothed martingale residuals versus Z will display the form of f. That is

$$E(M_i) \approx cf(z_i)\,,$$

where c is a constant which depends on the amount of censoring. Since c simply scales the labeling of the y axis, it has no effect on the visual appearance of the plot. In many ways, this plot is similar in spirit to the y vs. x scatterplots used for ordinary, non-survival data; the censoring process forces us to use this modification.

The following example shows the creation of such a plot in S, using the PBC data set.

```
> fit.pbc0 <- coxph(Surv(futime, fustat) ~ 1,  data=pbc)
> rr <- resid(fit.pbc0)           #default is martingale residuals
> plot(pbc$age, rr, xlab="Age", ylab='Residual')
> lines(lowess(pbc$age, rr, iter=0), lty=2)
> title(main="PBC data", sub='Figure 1')
>
> plot(pbc$bili, rr, xlab="Bilirubin", ylab='Residual')
> lines(lowess(pbc$bili, rr, iter=0), lty=2)
```

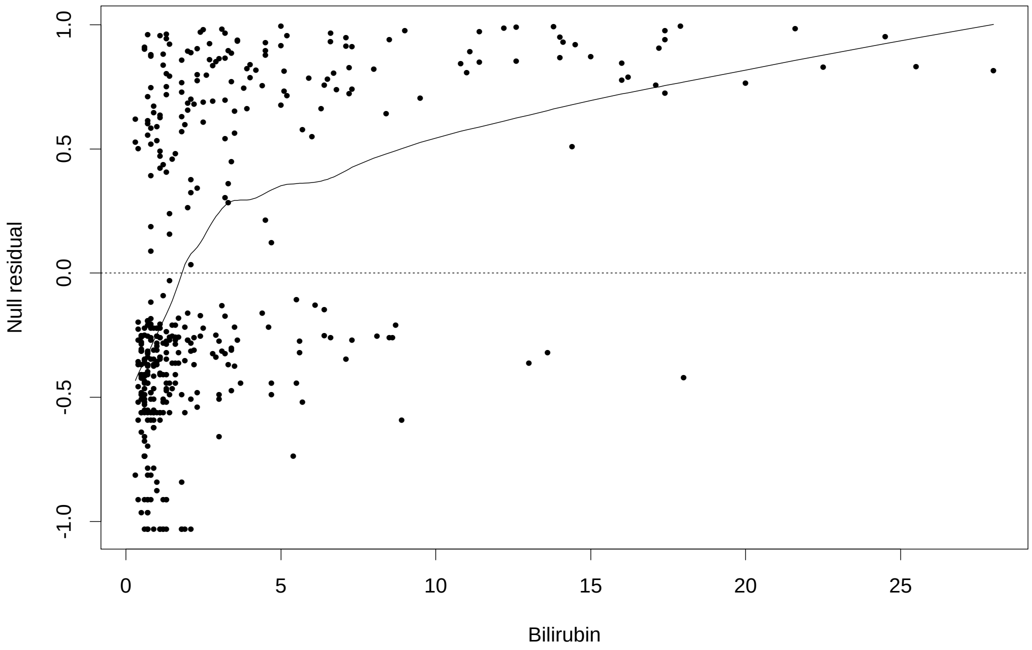

Figure 2: PBC Data, functional form for bilirubin

```
> title(main="PBC data", sub='Figure 2')
```

The `resid` function returns martingale residuals by default. The `iter=0` option of the lowess smoother causes it to skip its outlier detection phase, and is necessary because of the extreme skewness of the martingale residuals. (A similar problem can occur with logistic regression residuals when the proportion of postive responses is small). Figures 1 and 2 show that age is reasonably linear, but that bilirubin is certainly not so.

Creation of this same plot in SAS is more subtle, as `phreg` has no facility to fit a model with $\hat{\beta}$ forced to a given value; a "not converged" error message is invariably produced. However, a null model fit may be accomplished by using a single dummy variable as follows.

```
libname save 'sasdata';

data temp; set save.pbc;
    dummy = 0;

proc phreg data=temp;
    model futime*fustat(0)= dummy;
    output out=temp2 resmart=rr / order=data;
```

This method works well when the data are uncorrelated, but fails when correlations are present. The same failure occurs for ordinary scatter plots in uncensored data: if $y = 2x_1 + 0x_2$ and $\mathrm{cor}(x_1, x_2) = .9$, then a plot of y versus x_2 will show an unwanted relationship. There are other problems with the martingale residuals plot as well: the skewness of the data may force the fitted curve to occupy only a small region, and there are no clear methods for creating a confidence band.

4.2 Poisson approach

Grambsh, Therneau and Fleming [17] later extended the method to address these deficiencies. The basic idea is to use the residuals from a linear fit as the building block. Let M_i be the martingale residuals from a Cox model using all of the covariates, and E_i the expected number of events for each subject, based on the fitted model. (Since the residual satisfies $M_i = N_i - E_i$, the expected number of events is just the 0/1 status variable minus M_i). Two different plots are suggested. The first is based on the relationship

$$f \approx \log \left\{ \frac{\mathrm{smooth}(N)}{\mathrm{smooth}(E)} \right\} + \hat{\beta}_j Z_j + c$$

where c is an unknown additive constant. If all of the covariates are linear except for the jth, this should reveal the functional form of that covariate, Z_j. The approach performs well, but can experience numerical difficulties since $\mathrm{smooth}(E)$ may not always be > 0.

Their second idea, illustrated here, is to use the residual data as input to Poisson regression, taking advantage of modeling tools already available for that method. In S-Plus for instance this involves the gam function, which implements the Generalized Additive Models of Hastie and Tibshirani [20].

```
> fit.pbc <- coxph(Surv(futime, fustat) ~ age + edema + bili + protime
                    + albumin, data=pbc)
> print(fit.pbc)

           coef exp(coef) se(coef)     z       p
    age  0.0383      1.04  0.00806  4.75 2.0e-06
  edema  0.9351      2.55  0.28186  3.32 9.1e-04
   bili  0.1158      1.12  0.01302  8.90 0.0e+00
protime  0.2006      1.22  0.05661  3.54 3.9e-04
albumin -0.9682      0.38  0.20533 -4.72 2.4e-06

Likelihood ratio test=182  on 5 df, p=0  n=416
```

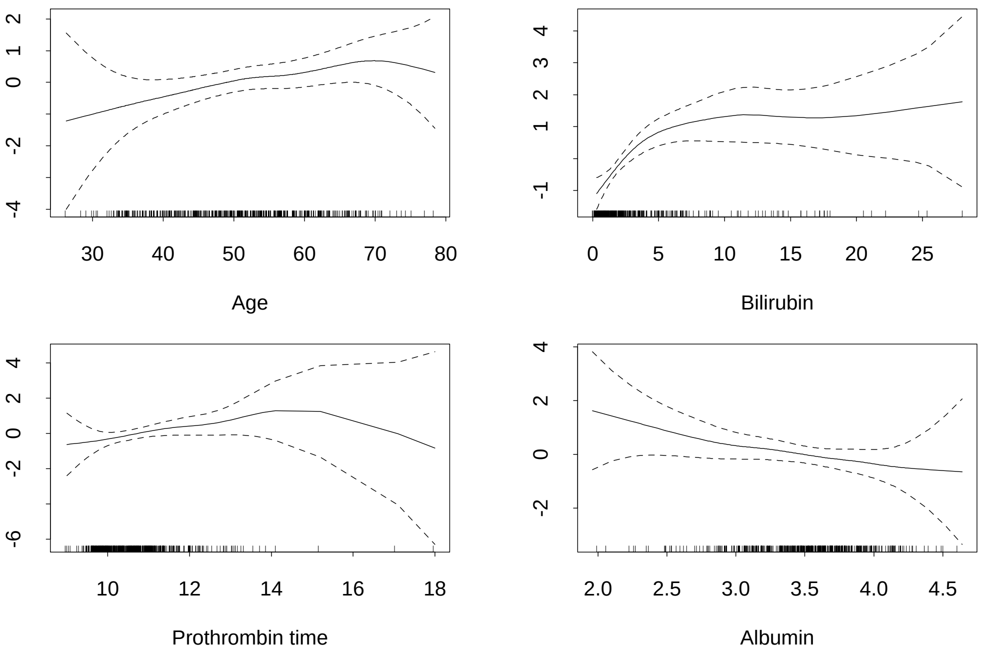

Figure 3: PBC Data, functional form using Poisson approach

```
(2 observations deleted due to missing)

> exp.pbc <- predict(fit.pbc, type="expected")
> xbeta   <- predict(fit.pbc, type="lp")       #the linear predictor
> gfit <- gam(fustat ~ s(age) + edema + s(bili) + s(protime) +
                    s(albumin) + offset(log(exp.pbc) - xbeta),
                data=pbc, family=poisson, na.action=na.omit)
> plot(gfit, se=T, rug=T)
```

For the PBC data set prior work had shown that age and edema were reasonably modeled as linear terms, and that bilirubin, prothrombin time and albumin levels fit well with a logarithmic transform. Hopefully, the plots will clearly reveal this trend.

In the gam function, the s(age) term asks for the fit of a smoothing spline in age with the default 4 degrees of freedom. The edema variable has only 3 values, 0, .5 and 1, and is fit as a linear term. The offset term includes both the log(expected) term usual to a Poisson model, but also the linear predictor $X'\hat{\beta}$. This causes the linear term, already fit by the Cox program, to be reflected in the plots.

The plots are shown in figure 3 (a linear plot for edema is also produced by the plot command, but is omitted). We can see the logarithmic form clearly for bilirubin. A transform does not appear necessary for protime or albumin. The `rug` option to the plot command produces the set of tick marks along the bottom of the plot, one at the location of each of the x-values for the data. One can see from the rug that the apparent downturn at the right extreme of the protime plot is based on only 2 data points.

A logical next step would be to replace bilirubin with its logarithm and repeat the process, however, we will jump directly to the final model. (Bilirubin is the dominant variable in the model, and any changes to its modeling are reflected in all the other plots).

```
> fit.pbc2 <- coxph(Surv(futime, fustat) ~ age + edema + log(bili) +
                                             log(protime) + log(albumin),
                                      data=pbc, method="breslow")
> print(fit.pbc2)

                coef exp(coef) se(coef)      z       p
         age  0.0396    1.0404  0.00767   5.16 2.4e-07
       edema  0.8946    2.4463  0.27165   3.29 9.9e-04
   log(bili)  0.8630    2.3703  0.08295  10.40 0.0e+00
log(protime)  2.3856   10.8654  0.76876   3.10 1.9e-03
log(albumin) -2.4966    0.0824  0.65280  -3.82 1.3e-04

Likelihood ratio test=231  on 5 df, p=0  n=416
 (2 observations deleted due to missing)

> exp.pbc <- predict(fit.pbc2, type="expected")
> lbili <- log(pbc$bili)
> lpro  <- log(pbc$protime)
> lalb  <- log(pbc$albumin)
> xbeta <- c(cbind(lbili, lpro, lalb) %*% fit.pbc2$coef[3:5])

> gfit <- gam(fustat ~s(lbili) + s(lpro) + s(lalb)
                             + offset(log(exp.pbc) - xbeta)
                         data=pbc, family=poisson, na.action=na.omit)
> plot(gfit, se=T, rug=T)
```

In this run we have also made the refinement of dropping out the linear terms for age and edema, both from the `gam` fit and from the compensating variable `xbeta`. Since we are satisfied with them as linear terms, and the `coxph` program has already modeled them as such, they can be excluded from the `gam` fit. The `gam` fit will be more stable since it has fewer terms. The final plots are shown in figure 4, and show that the transformations are quite satisfactory.

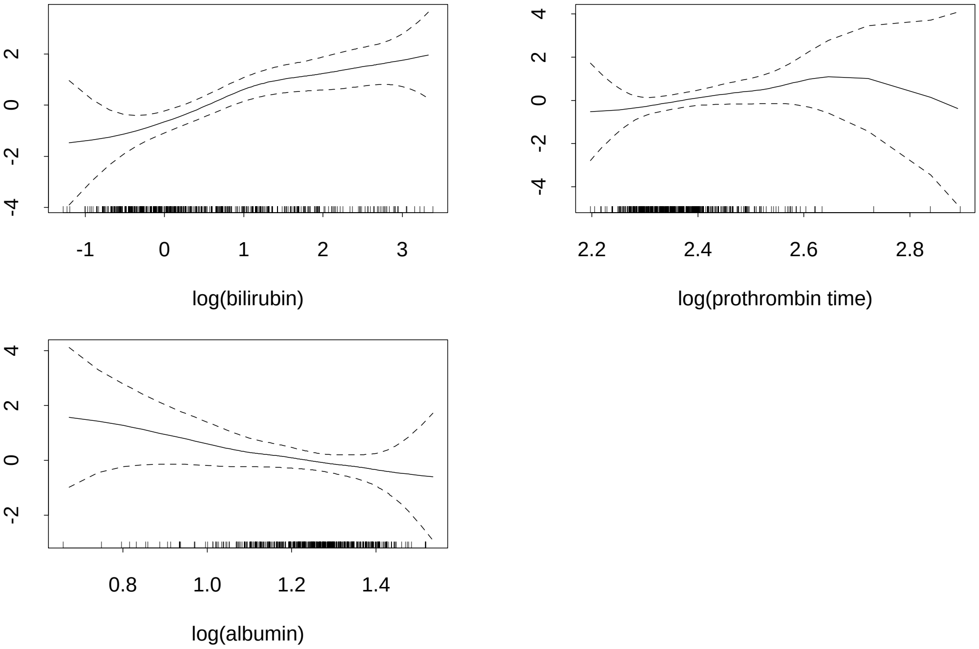

Figure 4: PBC Data, test of final functional form

With smaller data sets and/or a large number of variables, this method should be applied to one variable at a time rather than all-at-once, to avoid an excessive number of degrees of freedom in the gam model. Grambsch et al [17], however, show one constructed example with highly correlated predictors and nonlinear effects where one-at-a-time plots do not completely succeed.

For the SAS code, we cannot take advantage of a standard additive models procedure. Instead, for each predictor variable in turn a natural spline is fit using the genmod procedure, followed by a plot of the predicted values of the fit. The daspline macro creates a set of basis vectors age1, age2, age3, age4 which allows a spline to be fit with standard procedures.

```
proc phreg data=save.pbc outest=fit;
    model futime*fustat(0) = age edema bili protime albumin
                                    /ties=efron;
    output out=temp resmart=rr/ order=data;

data temp2; merge rr save.pbc;
    keep fustat rr expected age bili protime albumin;
    expected = fustat - rr;
```

```
data temp3; set fit;
    keep beta;
    beta = age;
data temp4; merge temp2 temp3;
    xbeta = age * beta;
    %daspline(age, 4);              * spline basis with 4 df ;
    off = log(expect) - xbeta;
proc genmod data=temp4 noprint;
    model fustat = age age1 age2 age3 / offset=off dist=poisson;
    make 'obstats' out=temp5;

*plot of x=age, y=xbeta, xbeta-2*std, xbeta+2*std;
data temp6;
    merge temp4 temp5;
    lower = xbeta - 2*std;
    upper = xbeta + 2*std;
    res   = xbeta + resdev;
proc sort; by age;
proc gplot;
    plot xbeta*age=1
         lower*age=2
         upper*age=3
         res*age  =4  /overlay vaxis=axis1 haxis=axis2;

    symbol1 i=join l=1;
    symbol2 i=join l=2;
    symbol3 i=join l=2;
    symbol4 i=none v=dot h= .1 cm;

    axis1 label=(r=0 a=90 "smooth(age), df=4");
    axis2 label=("age");

*repeat the temp3 through plot process for bili, then protime, etc.
```

4.3 Other methods

Another method to adjust the plots for possible correlation between the predictor variables is to borrow techniques from the linear models literature. Let M' be the martingale residual from a model omitting age. The *adjusted variable* plot uses M' on the y-axis, and the residual from a linear regression of age on the remaining predictors in the model on the x-axis. If the plot is linear, then age is presumed to enter linearly into the multivariate model. Other variants on this theme are the *partial residual* and *augmented partial residual* plots. See Chambers et al [9] for an explanation and examples of the methods.

Chen and Wang [7] discuss another method, *constructed variable* plots, which should be useful for detecting a power transform. They are based on the Taylor series expansion

$$x^{(\lambda)} \approx x + (\lambda - 1)(x \log x - x) + (\lambda - 2)\,,$$

where

$$x^{(lambda)} \equiv \begin{cases} (x^\lambda - 1)/\lambda & if \lambda \neq 0 \\ \log x & if \lambda = 0 \end{cases}$$

which suggests the use of $z = x \log x - x$ as the extra variable in an adjusted variable plot. The slope of a fitted line will suggest the appropriate power transform.

In the author's experience, none of these methods work as reliably as the Poisson regression based fits of the section above.

4.4 Time dependent covariates

The martingale residual for a subject with time-dependent covariates is well defined, although the computation requires more bookkeeping. (With a time-dependent covariate defined by programming statements, the SAS `phreg` procedure presently will not produce the martingale residual). However, it is not at all clear how to plot each observation — what should be used as the x-axis value?

If the time-dependent covariate is encoded using the counting process style of data, however, the above methods can be used. In this case each subject is represented as one or more observations, each consisting of a time interval, the status, and *fixed* covariate values over that interval. Both SAS and S-Plus return the martingale residual *per observation.* The total residual for a subject is then the sum of residuals, over the observations that represent his/her data.

One could explore functional form using the simpler method by plotting one point for each observation. Depending on their length of follow-up, different subjects may have different numbers of observations, however, and for a given subject, observations may encompass different intervals of time. This can introduce bias in the simple method by effectively giving different weights to subjects, e.g., a person with 10 observations (10 points on the scatter plot) will have a larger influence on the smooth than a subject with only one observation.

The Poisson based method, however, automatically provides the correct weighting through the expected values e_i and can be used without modification in such a data set. One exercise, not shown here, is to randomly divide some of the subjects in the PBC data set into two intervals. It is easy to show that the fitted Cox model and the Poisson regression estimate of functional form are unchanged by this manipulation.

5 Testing proportional hazards

5.1 Time dependent coefficients

Many possible alternatives to proportional hazards exist. One easily expressed generalization is provided by models with a *time-dependent coefficient*

$$\lambda(t) = \lambda_0(t) \exp[\beta(t) Z] \,.$$

This is not at all the same as a time-dependent *covariate* $Z(t)$. The proportional hazards model, for a given covariate j, corresponds to the restriction $\beta_j(t) = \beta$, i.e., that a plot of $\beta_j(t)$ versus time will be a horizontal line.

Let $V(\beta, t)$ be the covariance of Z at time t, defined analogously to the running mean $\bar{Z}$, whose j, j' element is estimated as

$$V_{jj'}(\beta, t) = \frac{\sum_i Y_i(t) r_i(t) [Z_{ij}(t) - \bar{Z}_j(t)][Z_{ij'}(t) - \bar{Z}'_j(t)]}{\sum_i Y_i(t) r_i(t)} \,,$$

let s_k be the Schoenfeld residual for the *kth* death in the study, and s^*_k be the rescaled Schoenfeld residual $V^{-1}(\beta, t_k) s_k$.

Grambsch and Therneau [16] show that if $\hat{\beta}$ is the coefficient from an ordinary fit of the Cox model, then

$$E(s^*_k + \hat{\beta}) \approx \beta(t_k) \,.$$

This suggests a plot of $s^* + \hat{\beta}$ versus time or some function of time $g(t)$ as a method for visualizing the extent of non-proportional hazards. A line can be fit to the plot followed by a test for zero slope, a non-zero slope is evidence against proportional hazards. If T is the test statistic for zero slope, it is clear that different choices of the time-scale g lead to different tests for model misspecification.

1. If $g(t)$ is a specified function of time, then T is a score test for the addition of the time-dependent variable $g(t) * Z$ to the model, a test initially suggested by Cox [8]. Chappell [6] describes the relationship between this test and the test of Gill and Schumacher [15].

2. If g is piecewise constant on non-overlapping time intervals with the intervals and constants chosen in advance, T is the score test proposed by O'Quigley and Pessione [40], which generalizes and extends goodness of fit tests proposed by Schoenfeld [45] and Moreau, O'Quigley and Mesbah [36]. As the authors point out, this test has the disadvantage that the investigator must choose a partition of the time axis, but they suggest guidelines for doing so.

3. If $g(t) = \overline{N}(t-)$ then T is the covariance between the scaled Schoenfeld residual and the rank of the event times. The resulting test is similar to one proposed by Harrell [18], who uses the correlation between the unscaled residuals and rank of the event times. This test is familiar to users of the (now discontinued) SAS phglm procedure.

4. Lin [28] suggests comparing $\hat{\beta}$ to the solution $\hat{\beta}_g$ of a weighted Cox estimating equation
$$\sum_i \int g(t)\,[Z_i(t) - \bar{Z}(t)] dN_i(t) = 0$$
with $g(t)$ one of the scalar weight functions commonly chosen for weighted log rank tests, and he showed that asymptotically $\hat{\beta} - \hat{\beta}_g$ is multivariate normal with mean$\sim$0 and a variance matrix derived from martingale counting process theory. If the estimator $\hat{\beta}_g$ were based on a one-step Newton-Raphson algorithm starting from $\hat{\beta}$, his test would be identical to T. Lin suggested a monotone weight function such as $\widehat{F}(t)$, the left-continuous version of the Kaplan-Meier estimator for the survivor function of the entire data set, to detect monotone departures from proportionality and a quadratic function such as $\widehat{F}(t)\{1 - \widehat{F}(t)\}$ for non-monotone trends.

5. Nagelkerke, Oosting and Hart [38] suggest using the serial correlation of the Schoenfeld residuals for a univariate predictor, or for multivariate covariates, the correlation of a weighted sum, $a's$. The authors standardize by using a permutational approach to estimate the variance. They suggest $a = \hat{\beta}$ as a natural choice for the weights, followed by examination of individual covariates if the test is significant. This is equivalent to using the lagged residuals as $g(t)$.

The key point is that each of the above tests can be directly visualized as a simple trend test applied to the plot of $g(t)$ versus the scaled residuals.

In practice, the individual estimates of V at each death time may be unstable, particularly near to the end of follow-up when the number of subjects in the risk set is less than the number of elements of V. For most data sets, the variance matrix of $Z(t)$ changes slowly, and is quite stable until the last few death times. Combining this with the observation that
$$\int_0^\infty V(\beta, t) d\overline{N}(t) = \mathcal{I}(\beta)$$
where $\mathcal{I}^{-1}$ is the Cox model's usual estimate for the variance of $\hat{\beta}$, suggests the use of the average value $\overline{V} = \mathcal{I}/d$, where $d = \overline{N}(\infty)$ is the total number of deaths. With this substitution, let $t_k, k = 1, \ldots d$ be the individual death times, $g_k = g(t_k)$ the

chosen transformation, $\bar{g} = \sum g_k/d$ the average of the transformed time values, and S the matrix of unscaled residuals. The tests are based on a slope parameter

$$\theta = D^{-1}[S'(g - \bar{g})]$$

with variance

$$D^{-1} = d\mathcal{I}^{-1} / \sum (g - \bar{g})^2 .$$

A global test of proportional hazards, over all p covariates is

$$T = \frac{(g - \bar{g})' S \mathcal{I}^{-1} S'(g - \bar{g})}{\sum (g_k - \bar{g})^2/d}$$

The test for an individual covariate Z_j is

$$T_j = \frac{\sum (g_k - \bar{g}) s^*_{kj}}{\sqrt{\mathcal{I}^{-1}_{jj} \sum (g_k - \bar{g})^2/d}} \tag{5}$$

Since the Schoenfeld residuals sum to zero, the above equation is the usual test of a correlation coefficient, with $\mathcal{I}$ as an estimator for the variance of s^*.

To aid in detecting the possible form of departure, a smooth curve with confidence bands is added to the plot. For both the S-Plus and SAS functions this has been done using a spline fit. Let X be the matrix of basis vectors for the spline fit of the scaled residual on $g(t)$ and B the same spline functions, but evaluated at the plotting points. (B will usually be based on 30-40 points evenly spread over the range of $g(t)$). The plotted values of the spline curve will be

$$\hat{y} = 1\hat{\beta}' + B(X'X)^{-1}X'R \equiv 1\hat{\beta}' + HR$$

where $H = B(X'X)^{-1}X'$ is the projection matrix. The variance matrix for the jth variable is

$$S_j = \mathcal{I}^{-1}_{jj}\{dHH' + (J - HJH')\}$$

where J is a matrix of 1's. For most smoothers, smooth(constant) = constant so that $HJ = J$ and the second and third terms cancel. The resultant formula is equivalent to the standard linear model's formula for a standard error for the predicted values, with the exception that $d\mathcal{I}^{-1}_{jj}$ replaces the usual estimator of σ^2. Confidence intervals can be formed by standard linear model calculations, e.g., Sheffé intervals using the rank of S_j for simultaneous confidence bands or simple z-intervals for pointwise estimates.

If the residuals are used as input to a "standard" scatterplot smoother, the resultant confidence bands will be based on $\hat{\sigma}^2 HH'$ where $\hat{\sigma}$ is based on the sum of squared residuals from the smooth. We have not done any theoretical investigation of this estimator, but note that

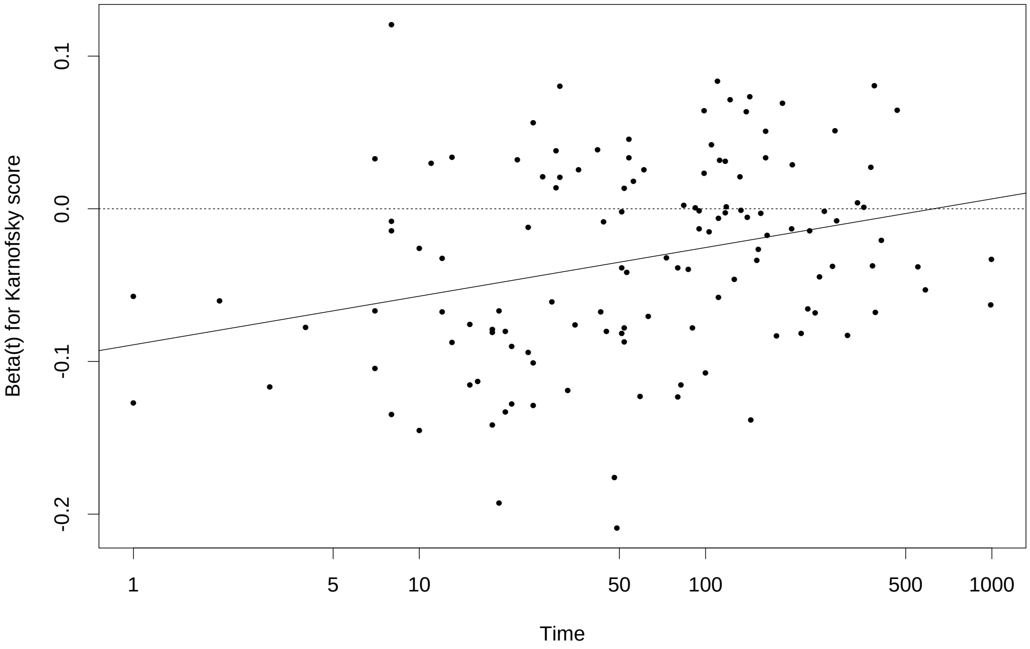

Figure 5: Veteran Data, tests of PH

- based on a *very small* number of empirical cases, the resultant bands tend to be somewhat too large (anticonservative),
- the *shape* of the bands will be correct, however.

5.2 Vetarans Administration data

As an example consider the Veterans Administration lung cancer data as found in Kalbfleisch and Prentice [22], pp. 223–224, from a clinical trial of 137 male patients with advanced inoperable lung cancer. The end point was time to death and there were six covariates measured at randomization: cell type (squamous cell, large cell, small cell, and adenocarcinoma), Karnofsky performance status, time in months from diagnosis, age in years, prior therapy (yes/no) and therapy (test chemotherapy versus standard). Lin's test [28] comparing the Cox model $\widehat{\beta}$ to a weighted estimate with the Peto-Prentice weight function is highly significant ($p = 0.00002$), suggestive of non-proportionality. Figure 5 shows the scaled Schoenfeld residuals for the most significant predictor, Karnofsky score, along with a fitted least squares line. As the results below show, the Grambsch-Therneau test for slope=0 is highly significant.

```
> fit.vet <- coxph(Surv(futime, status) ~ rx + celltype + karno +
                     months + age + prior.rx, data=veteran)
> print(fit.vet)
                      coef exp(coef) se(coef)     z      p
              rx  0.319242     1.376  0.20949  1.5239   .13
   celltypelarge -0.799691     0.449  0.30305 -2.6388   .008
celltypesmallcel -0.328601     0.720  0.27632 -1.1892   .23
celltypesquamous -1.236709     0.290  0.30491 -4.0560   .0005
           karno -0.032886     0.968  0.00553 -5.9471  <.0001
          months -0.000269     1.000  0.00914 -0.0295   .98
             age -0.009646     0.990  0.00932 -1.0346   .30
        prior.rx  0.084866     1.089  0.23312  0.3640   .72

Likelihood ratio test=62.7  on 8 df, p=1.39e-10  n= 136

> zph.vet <- cox.zph(fit.vet, transform='log')

> for (i in 1:4) {
     plot(zph.vet, var=i)
     abline(0,0, lty=3)
     title(main="Veteran data, test for PH")
     }
```

Karnofsky score and cell-type were the only significant predictors in the Cox model. Smoothed scaled Schoenfeld residuals plots for these predictors are shown in figure 6 and provide a visual interpretation of the non-proportionality. Because the survival times have a long-tailed distribution, $\log(t_k)$ is used for the x-axis. Use of the Kaplan-Meier values for the x-axis (Peto-Prentice scores) does a slightly better job of equi-spacing the plot points, but the figure is quite similar. Smoothed curves are shown along with pointwise 90% confidence intervals. Plots for the remaining four predictors (not shown) did not suggest significant nonproportionality. Table 1 summarizes individual predictor test statistics, using (5) with log(event∼times) for $g(t)$. (It differs from the results in [16], whose table 1 is an amalgamation of the tests from individual univariate models.)

The impact of Karnofsky score clearly changes with time. Early, a low score is protective. However, the effect diminishes over time and is effectively zero by 100 days. Another way of interpreting this would be that a 3-4 month old Karnofsky score is no longer medically useful. The downturn at the right end of the plot is likely an artifact of small numbers and disappears if the last four points are excluded.

The effect of each cell type is less clearly marked. The plots suggest that the increased relative risk due to small cell or adenomatous as compared to large cell may not persist beyond 100 days and that the squamous cell type may be protective in long-term survivors (beyond 100 days) as compared to large cell.

The `transform` argument of the `cox.zph` function allows for any monotone time transform, but the identity, logarithm, rank, and Kaplan-Meier options are the most commonly used. The `%schoen` macro only supports the four common options. The default is the Kaplan-Meier scale, and is based on the following rationale: if the transform g were such as to push most of the data to one end of the plot, leaving only 1 or 2 isolated points on the other extreme, then the test for slope=0 would be dominated by those extreme points. The K-M seems to do the most reliable job of spreading the data evenly over the horizontal range.

One problem with the K-M transform is that time may not vary smoothly over the range of the x axis of the plot. For final presentation, then, one of the simpler transformations may lead to plot that is easier to grasp.

5.3 Stratified models

Both S-Plus and SAS currently return the scaled Schoenfeld residuals based on an overall estimate of variance $V(t) \approx \mathcal{I}/d$. This average over the risk sets is appropriate if the variance matrix is fairly constant over those risk sets. One case where this may not be so is with stratified models. If there are stratum by covariate interactions, the averaging is almost certainly unwise. Consider the following example: assume that the 0/1 variable `rx` contains the treatment arm and that `rx1` is defined as

$$\text{rx1} = \begin{cases} \text{rx} & \text{if center} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume `rx2, rx3` and `rx4` are defined similarly and that `center` has values 1, 2, 3 and 4 for four participating centers in the study. The model could be

```
coxph(Surv(time, status) ~ rx1 + rx2 + rx3 + rx4 + strata(center))
```

Clearly, the variable `rx1` is identically zero in stratum 2, it has variance 0 within that stratum, and hence those data "points" can contribute no information on the

Covariate	*Chi-square*	*d.f.*	*p*
Cell type	7.39	3	0.0604
Karnofsky score	11.68	1	0.0006
Months since diagnosis	1.67	1	0.1955
Age	6.58	1	0.0103
Prior therapy	3.90	1	0.0482
Treatment	0.05	1	0.8287
Global test	27.22	8	0.0006

Table 1: Tests for the Veterans Administration data with g = log(time)

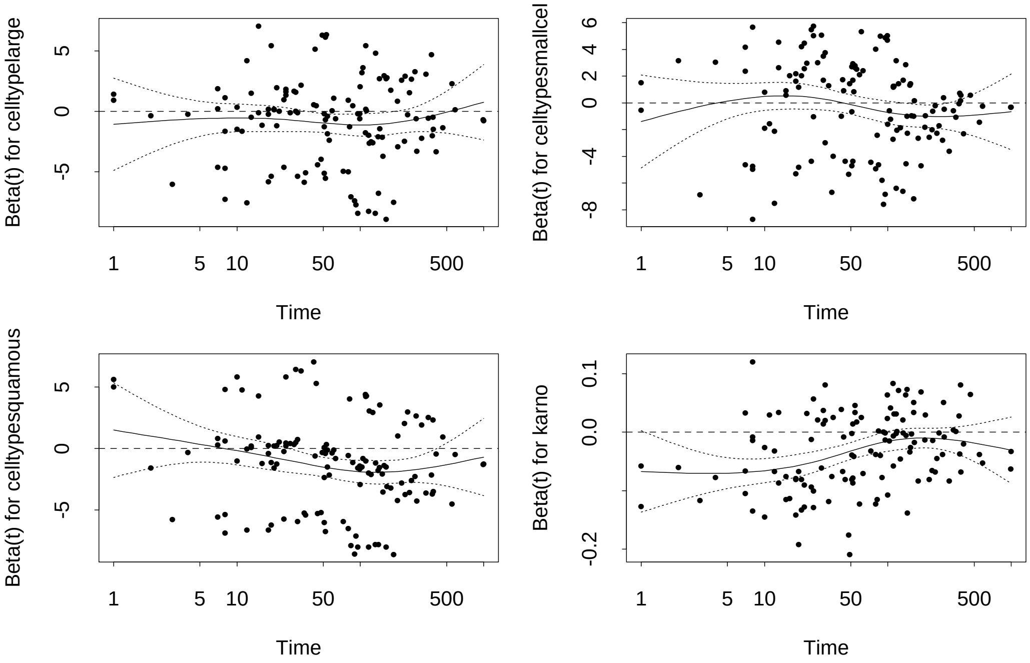

Figure 6: Veteran Data, test of PH for Karnofsky score

appropriateness of proportional hazards.

At present there is only a partial solution to this problem. First fit the overall model to the data. Then refit each stratum separately, using the `iter=0` and `initial` options to force the same coefficients as the overall fit. Since the variance will be summed only over the individual strata, this will produce scaled Schoenfeld residuals appropriate to the subsets.

6 Multiple events per subject

6.1 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 would be 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.

- A counting process approach, usually called the Andersen-Gill model [1]. Each subject is treated as a multi-event counting process with essentially independent increments. Any interrelation between events is modeled as one or more time-dependent covariates. This approach is simple, but the assumptions are strong and may be untenable.

- A marginal method as developed by Wei, Lin and Weissfeld [49]. They show the utility of the method for both a data set with multiple dissimilar outcomes (death and recurrence of cancer) and another with repeated outcomes (recurrence of bladder cancer).

- Frailty models, such as that described in Oakes [39]. He illustrates this 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% reduction in the variance of the treatment effect.

- A more ambitious plan is to model the subject's correlation directly within the Cox framework. Prentice and Cai [41] 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 section we focus on the AG and marginal models. This is partly due to the ready availability of software for this approach in both the S-Plus and SAS packages. As well, the method affords great flexibility in the formation of strata and risk sets, manipulation of the time scale, and has a well developed variance estimator. In each case the analysis is based on 3 steps:

- Decide on a model (issues such as strata, time dependent covariates, etc.) and structure the data set accordingly.

- Fit the data with an ordinary Cox model, ignoring the possible correlation.

- Replace the naive variance with a robust, corrected estimate.

In the sections below we will first deal with the third issue, then the first and second, and finally present a set of examples. 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, and an apparent treatment by time interaction. When applied to this data different models give apparently different answers.

6.2 Robust Variance

6.2.1 Approximate jackknife estimate

If one suspected that some element of the Cox model were misspecified, a natural correction would be to use the jackknife estimate of variance $(J-\bar{J})'(J-\bar{J})$, where J_{ij} is the change in $\hat{\beta}_j$ when observation i has been removed from the data set and $\bar{J} = 11'J/n$ is a matrix containing the column means of J. An natural approximation to the jackknife variance is $D'D$, where D is the approximate case influence introduced in section 3.3. (Remember that $1'D = 0$).

D can also be used to approximate a grouped jackknife, e.g., the sum of rows 1–3 of D approximates the change in $\hat{\beta}$ if observations 1–3 were removed 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 were formed from m groups of observations, of size $n_1, n_2, \ldots n_m$, with possible within-group correlation. Then one might form the collapsed $m \times p$ leverage matrix $\widetilde{D}$, where

$$\begin{aligned}
\widetilde{D}_{1j} &= \sum_{i=1}^{n_1} D_{ij} \\
\widetilde{D}_{2j} &= \sum_{i=1+n_1}^{n1+n2} D_{ij} \\
&\vdots
\end{aligned}$$

The kth row of $\widetilde{D}$ is an estimate of the leverage of the kth group, and $\widetilde{D}'\widetilde{D}$ approximates the grouped jackknife estimate of variance. The most common use of this estimate in our work will arise when there are multiple observations per subject. In this case the rows of D represent the *per observation* influence and those of $\widetilde{D}$ the *per subject* influence. Plots of both of these are useful in their own right for checking a fitted model.

6.2.2 Survey sampling method

A Cox model that includes case weights has been considered by Binder [4] in the context of survey data. If w_i are the weights, then the modified score statistic is

$$U(\beta) = \sum_{i=1}^{n} w_i u_i(\beta)\,, \tag{6}$$

where the u_i are the score residuals of section 3.3. (It can also be defined as a weighted sum of Schoenfeld residuals; the equation's solution is the same). Other formulae change in the obvious way, e.g., the weighted mean $\bar{Z}$ is changed to include both the risk weights r and the external weights w. The information matrix can be written as $\mathcal{I} = \sum \delta_i w_i v_i$, where δ_i is the censoring variable and v_i is a weighted covariance matrix. Again, the definition of v_i changes in the obvious way from equation (3). If all of the weights are integers, then for the Breslow approximation this reduces to ordinary case weights, i.e., the solution is identical to what one would obtain by replicating each observation w_i times. (With the Efron approximation or the exact partial likelihood approximation, of course, replication of a subject would result in the computer algorithm applying a correction for ties.) Residuals from the fit are such that the sum of weighted residuals equals 0.

If the observations are independent, the robust sandwich estimator is $D'WD$, where W is a diagonal matrix of the weights. When there is clustering, a standard survey sampling estimator based on linearization followed by a between cluster estimate has been used for many years, see for example Cochran [10]. This standard estimate reduces, in the Cox model case, to $\widetilde{D}'W\widetilde{D}$ [4].

6.2.3 Sandwich estimates

A rigorous motivation is based on the *sandwich estimate* of variance

$$V = ABA'$$

where $A^{-1} = \mathcal{I}$ is the usual information matrix, and B is a correction term. The genesis of this formula can be found in Huber [21], who discusses the behavior of any solution to an estimating equation

$$\sum_{i=1}^{n} \psi(x_i, \hat{\beta}) = 0\,.$$

Of particular interest is the case of a maximum likelihood estimate based on distribution f (so that $\psi = \partial \log f / \partial\beta$), when in fact the data are observations from distribution g. Then, under appropriate conditions, $\hat{\beta}$ is asymptotically normal with mean β and covariance $V = ABA'$, where

$$A = \left(\frac{\partial EU(\beta)}{\partial \beta} \right)^{-1}$$

and B is the covariance matrix for $U = \sum \psi(x_i, \beta)$.

As a simple example consider generalized linear models. McCullagh and Nelder [34] maintain that overdispersion "is the norm in practice and nominal dispersion the exception." To account for overdispersion they recommend inflating the nominal covariance matrix of the regression coefficients $A = (X'WX)^{-1}$ by a factor

$$c = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V_i} / (n - p)\,,$$

where V_i is the nominal variance. Smith and Heitjan [46] show that AB may be regarded as a multivariate version of this variance adjustment factor, and that c and AB may be interpreted as the average ratio of actual variance $(y_i - \mu_i)^2$ to nominal variance V_i. By premultiplying by AB, each element of the nominal variance-covariance matrix A is adjusted differentially for departures from nominal dispersion.

For the Cox model the matrix A is the usual variance-covariance matrix $\mathcal{I}^{-1}$ and $B = U'U$ where U is the matrix of score residuals. (B is the empirical covariance of the score vector). Thus $V = ABA = D'D$. If there is correlation, B is based on the collapsed score residuals $\tilde{U}$ and then $V = \tilde{D}'\tilde{D}$.

6.2.4 Relation to other methods

These estimates are also familiar from other contexts, although the general form $D'D$ has not always been emphasised. Using the same method of derivation as Cain and Lange [5], the results for a linear model are $U_{ij} = X_{ij}(y_i - \hat{y}_i)$, $D = U(X'X)^{-1}$ and $D'D$ is the robust variance estimate proposed by White [50, 51] for linear models with heteroscedasticity or other model violations.

For a generalized linear model with log-likelihood function $l(\beta)$

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

and $\tilde{D}'\tilde{D}$ is the *working independence* estimate of variance proposed by Liang and Zeger [25] for generalized estimating equation (GEE) models.

Lipsitz, Laird and Harrington [27] use the Six Cities dataset to compare several estimators of variance for logistic regression with correlated data: the usual MLE estimate, the actual jackknife variance estimate, the approximate jackknife $D'D$, and a more refined one-step approximation that corrects as well for changes in the information matrix $\mathcal{I}$ due to deletion of an observation. (To derive the refined formula, frame logistic regression as an iteratively reweighted least squares problem and then apply the exact jackknife formula for a linear model to the last step of the iteration). Using a sample size of 30 and 2 covariates plus an intercept the most accurate confidence interval coverage was given by $D'D$. The refined approximation did less well and the actual jackknife did poorly; it appears to be overly sensitive to individual data points. At $n = 60$ the two approximate methods were nearly identical, with the jackknife still somewhat inferior.

Reid and Crépeau [44] derive D for the Cox model, and apply it to measure the leverage of individual cases. In an appendix they mention the possible use of $D'D$ as a variance estimate but do not pursue it.

Lin and Wei [29] show the applicability of Huber's work to the Cox partial likelihood, and derive the ordinary Huber sandwich estimate $V = \mathcal{I}^{-1}(U'U)\mathcal{I}^{-1} = D'D$. They also discuss situations in which this is estimate is preferable, including the important cases of omitted covariates and incorrect functional form for the covariate.

Lee, Wei and Amato [24] consider highly stratified data sets which arise from inter observation correlation. As an example they use paired eye data on visual loss due to diabetic retinopathy, where photocoagulation was randomly assigned to one eye of each patient. There are $n/2 = 1742$ clusters (patients) with 2 observations per cluster. Treating each pair of eyes as a cluster, they derive the modified sandwich estimate $V = \widetilde{D}'\widetilde{D}$, where $\widetilde{D}$ is obtained by summing D over each individual's pair of eyes. A subject with only one studied eye would have one (identical) row of data in both D and $\widetilde{D}$.

Assuming a data set eyes with variables subject.id, time, status (0=censored, 1=failure) and treatment, and with two observations (rows of data) per subject, one S-Plus program to perform the analysis is

```
> fit    <- coxph(Surv(time, status) ~ treatment, data=eyes)
> Dtilde <- residuals(fit, type='dfbeta', collapse=subject.id)
> newvar <- t(Dtilde) %*% Dtilde
```

The first statement fits an ordinary Cox model, without regard to the clustering variable subject.id. The second statement retrieves a copy of the collapsed score residuals matrix $\widetilde{D}$, and the third forms the modified sandwich estimate of variance. (In S, %*% is matrix multiplication, t is the transpose function, and Surv is a 'packaging' function that allows both time and status to be part of the left hand side of a model formula). A second example given in Lee, Wei and Amato concerns a litter-matched experiment. In this case the number of rats/litter may vary. The resulting estimator is shown to be much more efficient than an analysis stratified by cluster.

Wei, Lin and Weissfeld [49] consider multivariate survival times, an example being the measurement of both time to progression of disease and time to death for a group of cancer patients. The S-Plus code to perform their suggested analysis is similar to the above. The data set again contains $2n$ observations, time and status variables, subject id, and covariates. It also contains an indicator variable `etype` to distinguish the event type, progression vs. survival. The suggested model is stratified on event type, and includes all strata×covariate interaction terms.

Lin and Wei [30] use the difference between $\mathcal{I}^{-1}$ and $D'D$ as a goodness of fit test for the Cox model. The variance estimate is quite complex, however, and I am not aware of any implementations of the method.

6.3 Setting up the problem

The prior section has dealt with the estimation of variance for a multi-event model. In this section we will discuss the options for setting up the data set. It turns out that the actual fitting of the model is extremely easy given these two steps.

One aspect of multiple event data sets is that there are a number of choices to be made 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:

- 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 stratum.
- The time scale is almost invariably time since entry to the study.
- Time dependent covariates usually reflect time dependent data such as repeated lab tests. Strata by covariate interactions, i.e., separate coefficients within each stratum for some covariate, are occasionally used.
- The counting process form may be used for a time dependent covariate, but normally the data set will consist 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 would be the times to death and progression for a set of cancer patients. For unordered outcomes setup of the data 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, e.g., 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 the usual variance estimator may still be used despite the correlation, see Lunn and McNeil [33]. The authors also compare models which stratify on the event type to ones which use event type as a covariate.

For ordered outcomes, i.e., multiple events of the same type, several suggestions have been offered. The most common approaches are the independent increment, marginal, or conditional models. All three are "marginal" regression models in that $\hat\beta$ is determined from a fit that ignores the correlation followed by a corrected variance $\widetilde{D}'\widetilde{D}$, but differ considerably in their creation of the risk sets. Lin [31] gives a detailed comparison of the three approaches on four different data sets.

6.3.1 Independent Increment model

This is usually referred to as the "Andersen-Gill" formulation, although we prefer this title. It is the simplest method to visualize and set up, but makes the strongest assumptions. It is closest in spirit to Poisson regression, and can in fact be accurately approximated with Poisson regression software in the same manner that Laird and Olivier [23] 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 1 or two observations (depending on whether there was additional followup experience after the first event), and 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 they are used, would be based on the same considerations as for an ordinary single event model.

The key assumption of the model is that of independent increments, i.e., that the multiple observations for a given subject are independent. If this is so then the three variance estimates $\mathcal{I}^{-1}$, $D'D$ and $\widetilde{D}'\widetilde{D}$ should all estimate the same quantity. (Nevertheless, it is wisest to use the per subject jackknife estimate $\widetilde{D}'\widetilde{D}$). A second assumption is that time-dependent covariates may be used to capture the changes, if any, between event types. For instance, let z be the time dependent covariate "number of prior events". A model might include both treatment, z and their interaction.

6.3.2 Marginal model

The marginal data model is used by Wei, Lin and Wiessfeld [49] in their analysis of bladder cancer data, and is sometimes referred to as the "WLW" method. For this method, each event or event type is modeled as a separate stratum. Within each stratum, the data used is 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. Since all the time intervals start at zero the model can be fit without recourse to the counting process style of input.

In the WLW paper, all strata by covariate interaction terms were included in the model. In this case the individual coefficients can be obtained (and were) 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 the combined variance was estimated as

$$\begin{pmatrix} D_1'D_1 & D_1'D_2 & D_1'D_3 & D_1'D_4 \\ D_2'D_1 & D_2'D_2 & D_2'D_3 & D_2'D_4 \\ D_3'D_1 & D_3'D_2 & D_3'D_3 & D_3'D_4 \\ D_4'D_1 & D_4'D_2 & D_4'D_3 & D_4'D_4 \end{pmatrix}$$

where D_1 is the matrix of dfbeta residuals from the first fit, D_2 that from the second and etc. This is algebraically equivalent to $\hat\beta$ and $\widetilde{D}'\widetilde{D}$ from a combined fit over all four strata, where the combined model includes all covariate by strata interaction terms. Using a global fit rather than separate fits for each event type has some practical advantages:

- It is easier to code, particularly when the number of events per subject is large.
- Other models can be encompassed, in particular one need not include all of the strata by covariate interaction terms.
- There need not be the same number of events for each subject. The method for building up a joint variance matrix requires that all of the leverage matrices be of the same dimension, this would be violated if information on one of the failure types was not collected for some subjects.

6.3.3 Conditional model

As with the marginal model, each event or event type is assigned to a different stratum. However, the risk interval for later events does not start at zero. 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 would be 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 strata 2 this subject would be at risk from time 100 to 185. Oakes [39] argues strongly for the conditional approach, and states that the marginal method will be inefficient.

6.3.4 Comparison

Assume a subject with events at t_1, t_2, and t_3, with no further followup after time t_3. In all three formulations, he will be represented in the data set by 3 observations:

	Interval	Stratum
A-G	$(0, t_1]$	1
	(t_1, t_2)	1
	(t_2, t_3)	1
marginal	$(0, t_1]$	1
	$(0, t_2)$	2
	$(0, t_3)$	3
conditional	$(0, t_1]$	1
	(t_1, t_2)	2
	(t_2, t_3)	3

Another way to look at the difference is to consider the risk sets. Suppose that subject "Smith" has experienced his second event on day 32. Who are the subjects at risk? When looked at in this way, the three methods have a natural ordering.

A-G: All subjects who were under observation on day 32.

marginal: All subjects who were under observation on day 32,
and have not yet had a second event.

conditional: All subjects who were under observation on day 32,
have not yet had a second event,
and have experienced a first event.

For all three setups, it is also possible to use "time since last event" as the time scale. This is uncommon with the independent increment and marginal models, but has been explored for the conditional case by Prentice, Williams and Peterson [43].

6.4 Examples

Comparison of the three methods, and understanding of the variance estimator, is most easily done through a series of examples. The first three examples involve data where the events are unordered, with one fictitious and two real data sets. The remaining examples concern ordered events. The fictitious data in this latter case is very revealing, and serves as a guide for the analysis of the real data sets.

6.4.1 Doubled data

As the simplest example consider a data set with doubled observations, e.g., suppose that through some programming error each line of data has been entered twice. As a specific example we will use the 85 observations for time to first event of the bladder cancer data discussed in section 6.4.5. The variable `id` contains the subject identifier. The code below fits the data using S-Plus. A single addition `+ cluster(id)` to the model formula identifies the potential clustering, and causes an adjusted variance $\widetilde{D}'\widetilde{D}$ to be computed as well as the usual estimate.

```
fit1 <- coxph(Surv(futime, status) ~ rx + size + number + cluster(id),
                data=double)
print(fit1)

          coef    exp(coef)  se(coef) robust se     z       p
    rx -0.5299      0.589     0.2234     0.3174  -1.669 0.0950
  size  0.2403      1.272     0.0537     0.0748   3.214 0.0013
number  0.0701      1.073     0.0719     0.0893   0.785 0.4300
```

In this case, the uncorrected variance is exactly half of the correct value. The robust jackknife value captures this correction almost perfectly (the ratio above is 2.02). The coefficients are identical to those from the correct data.

A fit using SAS is more clumsy since the `phreg` does not directly handle the clustering. But it is instructive since it shows how the robust estimate can be obtained using only the dfbeta residuals; the technique should be useful in many other packages.

```
proc phreg data=double;
    model futime * status(0) = rx + size + number;
    output out=temp1 dfbeta= rx size number;
    id id;

proc sort data=temp1; by id;
proc means data=temp1 noprint;          *add up rows to get D tilde;
    by id;
    var rx size number;
    output out=temp2 sum=rx size number;

proc iml;                                    *compute matrix product;
    use temp2;
    read all varrx size number in x;
    v = x' * x;
    reset noname;
    vname = "rx", "size", "number";
    print, "Robust variance matrix",, v[colname=vname rowname=vname];
```

An extension of this example that merges the coefficients and robust variance into a single printout can be found in the SAS manual for phreg.

6.4.2 Diabetic Retinopathy study

This example is used in Lee, Wei and Amato [24] to motivate the robust estimator. Between 1972 and 1975 seventeen hundred forty-two patients were enrolled in the study to evaluate the efficacy of photocoagulation treatment for proliferative diabetic retinopathy; photocoagulation was randomly assigned to one eye of each study patient, with the other eye serving as an untreated control. A major goal was to assess whether treatment significantly delayed the onset of severe visual loss. Several other potentially important covariates such as age, gender, and length of diabetes were also recorded.

The set up for this problem is as easy as that for the doubled data. Since each eye is at risk of failure both before and after its companion has failed, the final data set consists of $2n$ observations each at risk from time 0 onwards. There is no obvious stratification variable; the risk set when an eye fails is *all* eyes that have not yet failed.

The authors show that the resulting estimate and its variance are much more efficient than a matched analysis, which places each pair of patients into a separate stratum. Interestingly, the robust variance estimate $\widetilde{D}'\widetilde{D}$ is smaller than the ordinary estimate in this case. Because the treatment is balanced within subjects there is an improvement analogous to a paired t-test.

6.4.3 UDCA in Patients With PBC

Primary biliary cirrhosis (PBC) is a chronic cholestatic liver disease characterized by progressive destruction of the bile ducts. PBC frequently progresses to cirrhosis, which may lead to death from liver failure unless liver transplant is offered — an extensive and costly procedure. Trials have been held for several promising agents, but an effective therapy remains elusive. Although progression of disease is inex-

	UDCA	Placebo
Death	6	10
Transplant	6	6
Drug toxicity	0	0
Voluntary withdrawal	11	18
Histologic progression	8	12
Development of varices	8	17
Development of ascites	1	5
Development of encephalopathy	3	1
Doubling of bilirubin	2	15
Worsening of symptoms	7	9

Table 2: Total numbers of events in the UDCA trial

orable the time course can be very long; many patients survive 10 or more years from their initial diagnosis before requiring a transplant.

A randomized double-blind trial of a new agent, ursodeoxycholic acid (UDCA), was conducted at the Mayo Clinic from 1988 to 1992 and enrolled 180 patients. The study is reported in Lindor et al [26]. For this analysis, we exclude 10 patients who had incomplete follow-up for some of the event types (they were the last 10 enrolled in the study). The endpoints of the study were pre-defined and are shown in table 2. Although nearly all of the comparisons favor UDCA, none are significant individually. The primary report was based on an analysis of time to the first event; 58/84 placebo and 34/86 UDCA patients have at least one event. An analysis that used all of the events would seem to be more complete, however, since it would be based on 93 placebo and 52 UDCA events, a gain in "information" of 57%.

The endpoints for a subject are all unique, i.e., no single patient had more than one instance of death, transplant, doubling of bilirubin, etc. Thus, time ordering of the events within an event type is not an issue. Three possible methods of analysis present themselves. The simplest of course is time to the first adverse event. Each patient has a single observation and correlation is not an issue. The second is a marginal analysis. The data set for the marginal method is essentially a concatenation of the 9 individual data sets that would be created for an analysis of time to death (censoring all other causes), time to transplant, time to withdrawal, etc. A third, but less compelling alternative is to use an Andersen-Gill model, treating all of the events as though they were a single type.

Summaries of the first two approaches are shown in Table 3. The covariates in the model are treatment and two of the stratification factors used in treatment assignment. Two outcomes are immediately obvious. First, the naive variance is an underestimate in the multiple event model — accounting for the within patient

		β	se(β)	robust se
Time to first event				
	treatment	-0.94	0.22	0.22
	bilirubin	0.74	0.19	0.20
	stage	-0.02	0.25	0.25
Marginal model				
	treatment	-0.80	0.17	0.23
	bilirubin	0.77	0.18	0.25
	stage	0.05	0.21	0.28

Table 3: Results of 2 models for the UDCA data

correlation is important. The second is that the robust variance is just as large as it was for first events only; the use of multiple events has added no information to the analysis! The standard error for treatment in the marginal model is actually somewhat higher.

A closer look at the data reveals the cause of the difficulty. Patients on the study returned for evaluation once a year, which is the point at which most of the outcomes were measured (excluding death and transplantation of course). One patient who had 5 events, for instance, has 4 of them recorded on 20 July 1990, followed by death on July 22. Similar outcomes are seen for many others. Figure 7 shows the event times for the 31 subjects with multiple adverse outcomes, with a circle marking each event. The data has been jittered slightly to avoid overlap. It appears that the use of multiple event types was useful in this study only to make the detection of "liver failure" more sensitive. Given that failure has occurred, the number of positive markers for failure was irrelevant. This example shows that multiple event analysis does not always lead to gains.

6.4.4 Hidden variable data

We now turn to the more complicated case of repeated events of the same type, where the ordering of the events must be addressed. In the marginal method, these are treated as though they were unordered but distinguishable events, i.e., first event, second event, etc are treated in the same way that survival, liver transplant, and worsening of symptoms were treated in the UDCA data set. Both the independent increments and conditional methods treat the data as time-ordered outcomes, differing only in their use of stratification (a difference which can cause major changes in the conclusions).

We will first illustrate the methods with a simple test case, one which shows

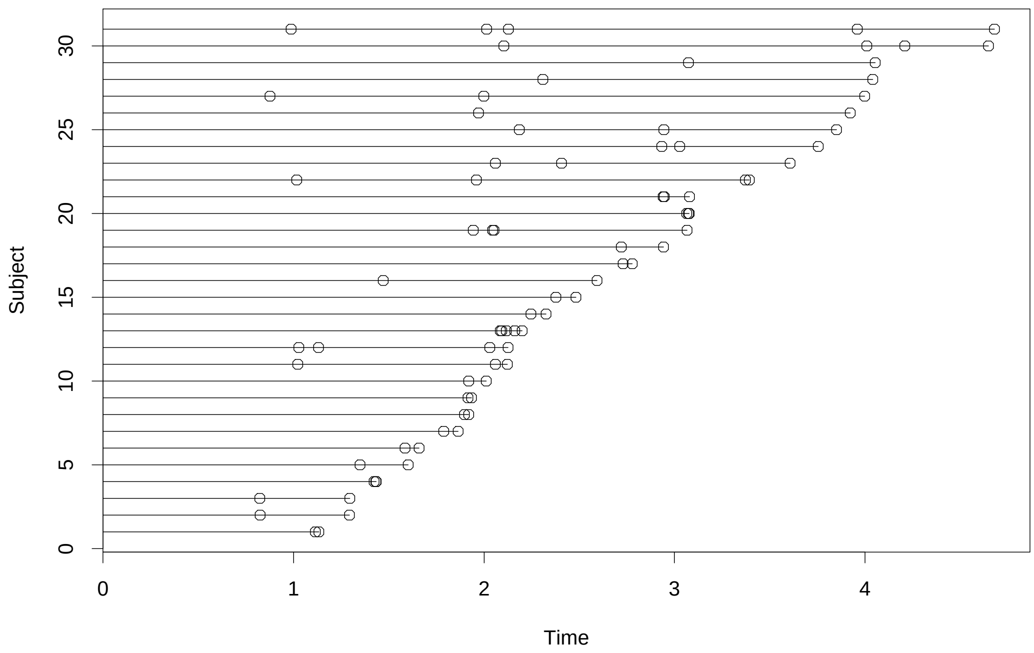

Figure 7: Multiple failure times for the UDCA data

that 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 1 year, which gave a mean number of events of 1.3. The sample size was 2000, which allows us to illustrate any biases in the estimates. For simplicity of presentation, the few subjects with more than 7 events were censored after their seventh. The number of events experienced was

	Number of events							
	0	1	2	3	4	5	6	7
Control	367	312	174	88	39	13	5	2
Treatment	680	250	50	14	6	0	0	0

To do the A-G and conditional analyses the data is set up in the counting process form. Assume that subject '10' is on treatment with a covariate value of .2, and has events on days 100 and 200, with follow up to day 365. The subject will be represented as 3 rows of data with variables

id	start	stop	status	enum	x_1	x_2
10	0	100	1	1	.2	1
10	100	200	1	2	.2	1
10	200	365	0	3	.2	1

For the marginal analysis we need a slightly different organization. Since at least one patient in the study had a seventh event, then all patients must be coded for that event type. The data rows for our fictional subject are

id	time	status	enum	x_1	x_2
10	100	1	1	.2	1
10	200	1	2	.2	1
10	365	0	3	.2	1
10	365	0	4	.2	1
10	365	0	5	.2	1
10	365	0	6	.2	1
10	365	0	7	.2	1

The modeling statements to fit the data for the three methods are similar. S-Plus statements for the A-G, conditional and marginal models, respectively, are

```
coxph(Surv(start, stop, status) ~ x1 + x2 + cluster(id),
                      data=data1)
coxph(Surv(start, stop, status) ~ x1 + x2 + cluster(id) +
                            strata(enum), data=data1)
coxph(Surv(time, status) ~ x1 + x2 + cluster(id) + strata(enum),
                      data=data2)
```

Consider testing for an overall treatment effect. An important point of comparison is the performance of the fit when x_1 is *not* included in the model. This corresponds, in real data sets, to those important covariates which are unmeasured or unknown to us. (The unmeasured covariate x_1 was purposely chosen to have a larger effect than the intervention.) In table 4 we see that the independent increments or AG model does fairly well in it's estimate of β. The standard error of the coefficient, however, is underestimated. When x_1 is included then the AG models is correctly specified, and both the coefficients and their standard errors are estimated without bias.

When x_1 is included then the conditional model is correctly specified as well, and the estimates are stable. When x_1 is unknown the conditional model seriously underestimates the treatment effect. This is due to a loss of balance in the unmeasured covariate. The mean level of x_1 for the first stratum (event number 1) is near 0 for both treatment and control. For stratum 2, however, the mean levels were 0.6

	without covariate	*with covariate*	
	β_2	β_2	β_1
Andersen-Gill			
coefficient	-0.92	-0.93	1.05
variance	.066, .084	.066, .066	.056, .056
marginal			
coefficient	-1.23	-1.60	1.82
variance	.066, .113	.069, .117	.063, .113
conditional			
coefficient	-0.67	-0.91	1.03
variance	.070, .068	.073, .069	.065 , .064

Table 4: Simple models, with both the naive and robust variance estimates

and 0.8, respectively: high risk patients are more likely to have an event, and since treatment is effective the treated patients must be, on average, of higher risk than controls to have had one. By strata 4, the baseline risk for the treatment arm is 40% greater than that for control. This is a serious problem and may preclude use of the conditional estimator in randomized trials.

Table 5 shows the estimated treatment effects within stratum for the marginal and conditional models, ignoring x_1 in the fit. (These can be obtained either by adding an interaction term `rx * strata(enum)` to the overall model, or by fitting each stratum separately). For stratum 1, time to first event, the conditional model is correct and reliable. The estimated treatment effect then steadily decreases for strata 2, 3 and 4. By stratum 5 there are only 6 treated patients, none of whom have an event, as compared to 13/39 control subjects. This leads to an infinite relative hazard (which is not, however, significantly different from zero by the likelihood ratio test). Strata 6 and 7 have no treatment subjects under the conditional setup.

The marginal model on the other hand overestimates the treatment effect, and inclusion of the covariate x_1 into the model does not rectify the problem. The per-stratum fits of table 5 show a steady growth of the estimated coefficient. (For stratum 1 the data set and fit are identical to the conditional model). The problem here is that the data for strata 2, 3, etc no longer obey the proportional hazards model. Strata 3, for instance, contains all 2000 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 for exponentials that the hazard ratio for strata k is $(\lambda_1/\lambda_2)^k$ at time 0 and then decays to an asymptotic value of λ_1/λ_2. Per strata checks of the proportional hazards assumption should be done when utilizing the marginal approach. Figure 8 shows the `cox.zph` plot with confidence bands for the first

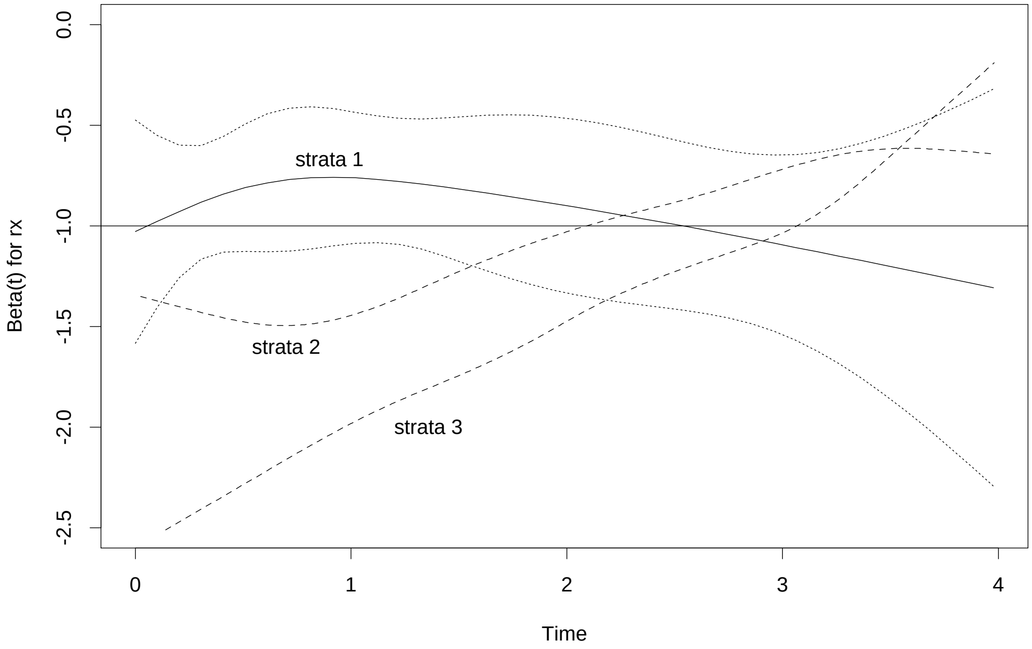

Figure 8: Tests of proportional hazards for the marginal model, strata 1–3

	rx1	rx2	rx3	rx4	rx5	rx6	rx7
marginal	-0.99	-1.7	-2.1	-2.3	-4.2	-4.2	-4.2
conditional	-0.99	-0.9	-0.6	-0.3	$-\infty$	NA	NA

Table 5: Models with interaction

stratum, with the curves for strata 2 and 3 overlaid. A horizontal line at $\beta = -1$ shows the true hazard ratio. The lack of proportional hazards for the latter strata is clear.

In summary, this example suggests

- The independent increment or Andersen-Gill model gives a nearly unbiased estimate of the treatment effect, even when an important covariate has been omitted. The naive estimate of variance may be too small, but the robust estimate $\widetilde{D}'\widetilde{D}$ corrects for this.

- The conditional model gives seriously biased estimates when an important covariate is omitted, due to swift loss of balance in the later strata.

- The marginal model may violate the proportional hazards assumption, even when the overall data set does not. Such violation can and should be tested for.

6.4.5 Bladder Cancer

The bladder cancer data is listed in Wei, Lin and Weissfeld [49] (WLW). The data set contains recurrence times in months for 86 subjects, each subject has between 0 and four recurrences. There may be follow-up beyond the last recurrence.

	Number of Recurrences				
	0	1	2	3	4
Number of Subjects	39	18	7	8	14
Follow-up after last event	38	17	5	6	12

One of the subjects (the one with no follow-up after the 0th recurrence in the table above) has no events and 0 months of follow-up. For simplicity, this subject was removed from the data set since he adds nothing to the likelihood. The covariates are initial size, initial number, and treatment group.

In order to accommodate all 3 analysis types, two data sets were created. Bladder2 contains 4 lines per subject with variables

id	subject id, 1 to 85
futime	follow-up or recurrence time
status	1=recurrence, 0=censoring
number	initial number
size	initial size
rx	treatment code, 1=placebo, 2=thiotepa
enum	event number

and is used for the marginal analysis. The `enum` variable will be 1 for the first recurrence, 2 for the second, etc. For a subject with recurrences at months 12 and 16 and further follow-up until month 18, the 4 observations will have values for `futime, status`, and `enum` of (12, 1, 1), (16, 1, 2), (18, 0, 3) and (18, 0, 4). The data set has 85*4 = 340 observations.

Note that the data set does not contain information on follow-up after the fourth event. These observations would be in stratum number 5, which has no events and so adds nothing to the analysis.

Data set `bladder1` is constructed in the AG style. In place of the `futime` variable there is a pair of variables `start, stop` which define the time interval of risk. Subjects with no recurrences will have one observation, those with 1 recurrence have 1 or 2 observations (depending on whether there is additional follow-up after the recurrence), and so on. Bladder1 has 190 observations.

The following lines replicate the suggested analysis in WLW:

```
> options(contrasts='contr.treatment')
> fit <-coxph(Surv(futime, status) ~ (rx + size + number) * strata(enum)
               + cluster(id), data=bladder2, method='breslow')
```

In order to match the WLW results, I have used the Breslow approximation for ties. The output of the fit is:

```
                  coef    exp(coef) se(coef)   robust se     p
           rx  -0.5176      0.596    0.3158      0.3075    0.092
         size   0.2360      1.266    0.0761      0.0721    0.001
       number   0.0679      1.070    0.1012      0.0853    0.430

     rx:enum=2  -0.1019      0.903    0.5043      0.3265    0.760
     rx:enum=3  -0.1823      0.833    0.5579      0.3916    0.640
     rx:enum=4  -0.1328      0.876    0.6581      0.4968    0.790

   size:enum=2  -0.0985      0.906    0.1193      0.1144    0.390
   size:enum=3  -0.0662      0.936    0.1298      0.1167    0.570
   size:enum=4   0.0930      1.098    0.1465      0.1175    0.430

 number:enum=2  -0.1440      0.866    0.1680      0.1119    0.200
 number:enum=3  -0.2792      0.756    0.2086      0.1511    0.065
 number:enum=4  -0.2708      0.763    0.2514      0.1856    0.140
```

The second three coefficients are treatment contrasts between stratum 1 and the others. WLW obtain their results by fitting each strata separately. This gives a coefficient for the treatment effect in strata 2, for instance, of (-0.5176) + (-0.1019) = -0.6195. Alternatively, specific dummy variables could have been created for use in the model equation to give the WLW coefficients directly, or one could create a contrast matrix `temp`

```
1 0 0  0 0 0  0 0 0  0 0 0
1 0 0  1 0 0  0 0 0  0 0 0
1 0 0  0 1 0  0 0 0  0 0 0
1 0 0  0 0 1  0 0 0  0 0 0
```

and compute `(temp %*% fit$coef)` for the coefficients and `(temp %*% fit$var %*% t(temp))` for the variance. Such manipulations should be familiar from linear models work.

WLW then suggest a combined treatment estimate, based on a variance weighted average of the four individual treatment estimates. A much easier approach is to just leave out the treatment by strata interaction:

		rx			size	
		naive	robust		naive	robust
	coef	se	se	coef	se	se
Andersen-Gill	-0.41	0.200	0.251	0.17	0.047	0.056
conditional	-0.33	0.216	0.205	0.12	0.051	0.048
marginal	-0.59	0.202	0.311	0.22	0.046	0.065

Table 6: Comparative fits for the bladder data

```
> fit2 <- coxph(Surv(futime, status) ~ rx + (number + size)*strata(enum)
                + cluster(id), data=bladder2, method='breslow')
```

Since none of the interaction terms were significant, we will compare the three fitting methods using a main effects model with only treatment and size. The Andersen-Gill, conditional, and marginal fits are obtained with the following S-Plus code.

```
fita <- coxph(Surv(time1, time2, status) ~ rx + size + cluster(id),
                         method='breslow', data=bladder1)
fitc <- coxph(Surv(time1, time2, status) ~ rx + size + cluster(id) +
             strata(enum), method='breslow', data=bladder1)
fitw <- coxph(Surv(futime, status) ~ rx + size + cluster(id) +
             strata(enum), method='breslow', data=bladder2)
```

The results are shown in table 6. The conditional model gives a smaller estimate than the Andersen-Gill and the marginal model a larger one, but all are well within one standard error of each other. Figure 9 shows a proportional hazards plot constructed in the same way as figure 8. Although there is some evidence of a right shift in the curves for stratum 1, 2, and 3, all of the variation is within the pointwise confidence bands for the first stratum's curve. For this data set, unlike the hidden covariate example, there does not seem to be significant violation of the proportional-hazards assumption within the later strata.

For comparison, we also show the SAS code to fit the WLW model. SAS versions prior to release 6.10 do not have a provision to return the leverage residuals D. The phlev macro, written by Eric Bergstralh and John Kosanke (Mayo Clinic) can be used to obtain the necessary result, however. First the interaction variables must be constructed, then the macro is called to obtain a phreg fit, an output data set containing the leverage residuals $\widetilde{D}$, and a data set containing the new variance matrix. The macro also produces an output listing similar to the S listing above.

```
data temp1; set bladder2;
        rx1 = rx * (enum=1);
```

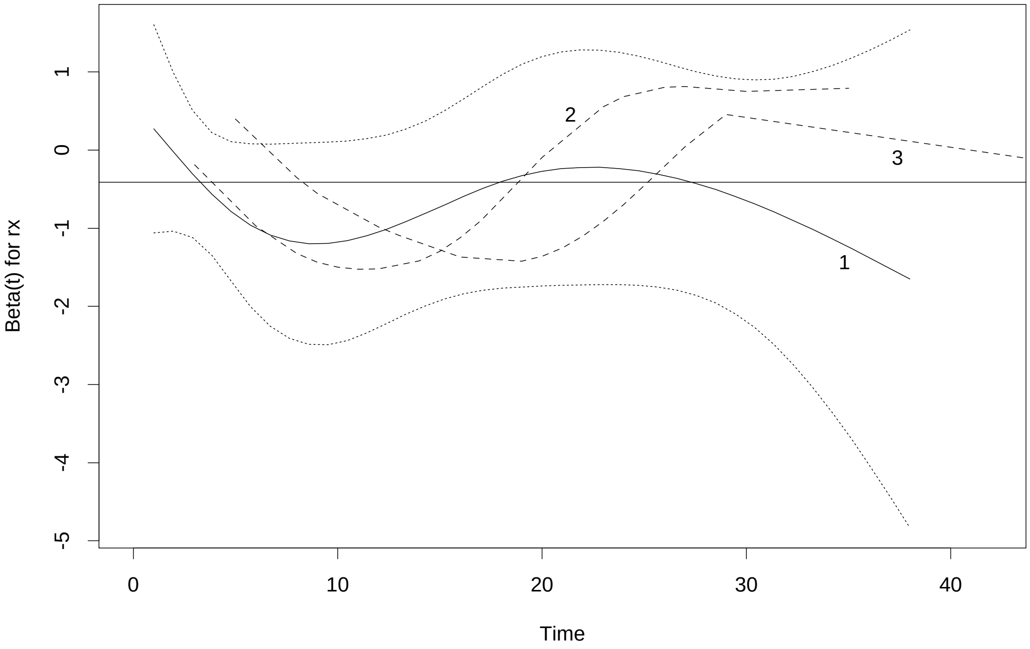

Figure 9: Tests of proportional hazards for the marginal model, strata 1–3

```
        rx2 = rx * (enum=2);
        rx3 = rx * (enum=3);
        rx4 = rx * (enum=4);
        number1 = number * (enum=1);
        number2 = number * (enum=2);
        number3 = number * (enum=3);
        number4 = number * (enum=4);
        size1 = size * (enum=1);
        size2 = size * (enum=2);
        size3 = size * (enum=3);
        size4 = size * (enum=4);

%let xx= rx1 rx2 rx3 rx4 number1 number2 number3 number4 size1 size2
                size3 size4;
title1 "Bladder Cancer Example";

%phlev(data=temp1, time=futime, event=status,
         xvars= &xx, strata=enum, id=id, collapse=Y);
```

The output of the macro comprises several pages. It includes the ordinary `phreg` output along with the following table. (The printed table includes more columns).

```
               Parameter                Robust      Robust
    Variable    Estimate        SE        SE      Chi-Square       P

    rx1         -0.51762    0.31576    0.30750       2.834      0.0923
    rx2         -0.61944    0.39318    0.36391       2.897      0.0887
    rx3         -0.69988    0.45994    0.41516       2.842      0.0918
    rx4         -0.65079    0.57744    0.48971       1.766      0.1839
    number1      0.06789    0.10125    0.08529       0.634      0.4260
    number2     -0.07612    0.13406    0.11812       0.415      0.5193
    number3     -0.21131    0.18240    0.17198       1.510      0.2192
    number4     -0.20317    0.23018    0.19106       1.131      0.2876
    size1        0.23599    0.07608    0.07208      10.720      0.0011
    size2        0.13756    0.09190    0.08690       2.506      0.1134
    size3        0.16984    0.10521    0.10356       2.690      0.1010
    size4        0.32880    0.12528    0.11382       8.345      0.0039
```

6.4.6 rIFN-g in Patients With Chronic Granulotomous Disease

Chronic granulotomous disease (CGD) 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 partially correct the metabolic defect in phagocytes, and for this reason it was hypothesised that it would 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. The resultant data set can be found in appendix D of Fleming and Harrington [14]. The primary endpoint of the study was the time to the first serious infection. However, data were collected on all serious infections until loss-to-followup, 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 numbers of infections were 56 and 20 in the placebo and treatment groups, respectively. Is a multiple events regression useful?

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 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

	β	se(β)	robust se
Time to first event	-1.09	0.34	0.34
A-G	-1.10	0.26	0.31
marginal	-1.34	0.27	0.36
conditional	-0.86	0.28	0.29

Table 7: Fits for the CGD data

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 A-G model.

The results of several models are shown in table 7. In the first model, time to first infection, the ordinary and robust variance estimates agree closely; a major disagreement would be evidence that some assumptions of the Cox model were violated. The Andersen-Gill model gives nearly an identical coefficient. If between subjects correlation is ignored then there is an apparent reduction in variance of 39%, from 0.112 to 0.068. Using the robust variance estimate $\widetilde{D}'\widetilde{D}$ the reduction is much smaller, only 13%. This suggests that including 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 ±.01 from these. If separated coefficients are fit to the first 3 strata or event numbers, the results for the marginal model are -1.10, -1.25 and -2.74, and for the conditional model they are -1.10, 0.11, and -1.28. (In the conditional setup, the treatment group has only 5 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 were to include significant covariates. The two most important factors, other than treatment, are age and enrollment center (the first 3 digits of the subject id). Table 8 shows the results for the treatment effect in a model that includes both age and center, the latter as a categorical variable with 13 levels. The results again parallel the hidden covariate data set. If center is entered as a stratification variable the results come even closer; the coefficients for the A-G and conditional models are -1.23 and -1.19, respectively.

6.4.7 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

	β	se(β)	robust se
Time to first event	-1.25	0.35	0.35
A-G	-1.16	0.26	0.30
marginal	-1.51	0.28	0.37
conditional	-1.00	0.29	0.29

Table 8: Fits for the CGD data, controlling for age and center

DNA thickens the mucus, which then cannot be cleared from the lung by the cilia. The accumulation leads to exacerbations of respiratory symptoms and progressive deterioration of lung function. More than 90 percent 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 Pulmozyme to placebo. 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, information on all exacerbations was collected for 169 days.

Table 9 shows the results on the number of exacerbations. Overall, 139/324 (43%) of the placebo and 104/321 (32%) of the rhDNase patients experienced an exacerbation during the follow-up period. A Cox proportional hazards model using the time to first exacerbations yeilds a hazard ratio of 0.69, with a 95% confidence interval of (.54, .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%) of the placebo and 39/104 (38%) of the treated patients who had a first exacerbation went on to experience a second. A multiple event Cox model can be used 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 suggest the use of either

- a marginal model with one stratum per event number, or
- a conditional model with one stratum per event number,
- an Andersen-Gill model with a time dependent covariate for event number

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

Table 9: Frequencies of exacerbations in rhDNase trial.

Arguing against this, however, is the fact that patients are entered into the study at various stages of disease. To attempt to control for $x_1 =$"number of events while on study" when $x_2 =$"number of events prior to entry" is unknown is not fruitful in the A-G model, especially when x_2 has both a larger mean and a larger variance than x_1.

Setting up the data sets for these models was more complicated than usual because of discontinuous intervals of risk. During an exacerbation, patients recieved IV antibiotics and were not considered at risk for a new event until seven exacerbation free days beyond the end of IV therapy. Figure 10 shows a plot of the time at which an event occurred versus the time not at risk for the event. (The horizontal coordinate mostly serves to spread out the points). There is no overall difference in time to return between the treatments, but the treated have less events and thus more total exposure time during the study.

Consider a single treated patient who had exacerbations at days 50 and 100 with durations of 10, and 15 days, respectively and a final follow-up at day 180. Two data sets were created to do the analysis. In the first, used for both the A-G and conditional analysis, this patient would appear as 3 observations with data values

```
time1   time2   status  rx      enum
  0       50       1     1       1
 60      100       1     1       2
115      180       0     1       3
```

For the marginal analysis, they will appear as 12 observations. (Since 1 person has 5 events, all subject appear in 5 strata.)

```
time1   time2   status  rx      enum
  0       50       1     1       1

  0       50       0     1       2
 60      100       1     1       2
```

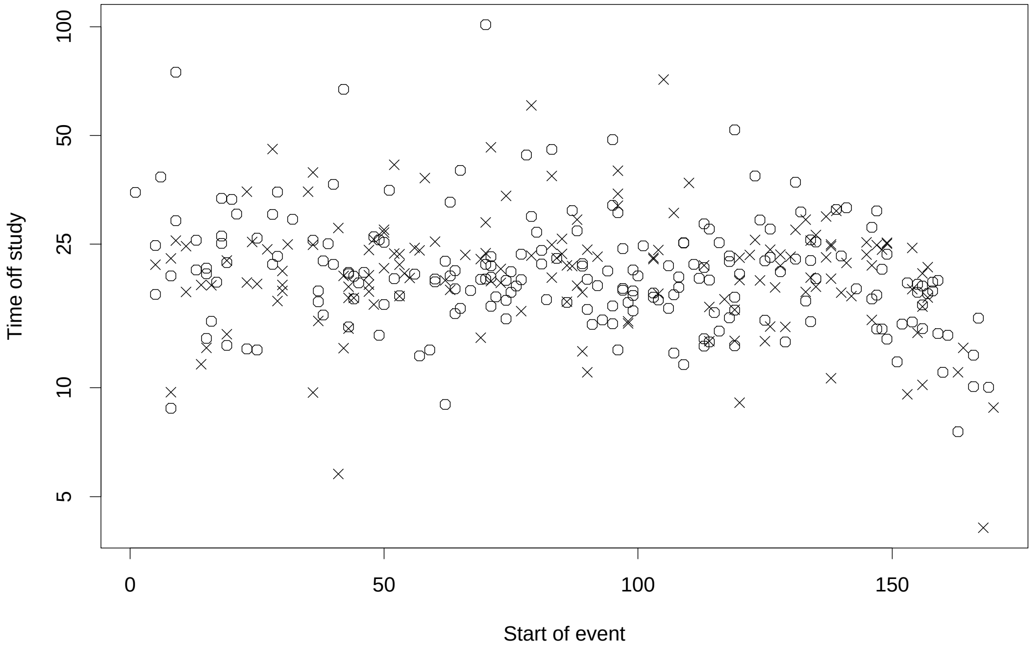

Figure 10: Time off study for dnase patients, o=placebo, x=treatment

```
       0        50           0       1          3
      60       100           0       1          3
     115       180           0       1          3

       0        50           0       1          4
      60       100           0       1          4
     115       180           0       1          4

       0        50           0       1          5
      60       100           0       1          5
     115       180           0       1          5
```

A further consideration is the very small numbers 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 stratum hazard estimates will be 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 stata 3–5. For the marginal model the strata may be preserved or

	β	se	robust se	p
First event	-0.365	0.13	0.13	0.005
Andersen-Gill	-0.287	0.11	0.13	0.029
Marginal	-0.336	0.11	0.15	0.021
Conditional	-0.220	0.11	0.11	0.043

Table 10: Simple fits to the dnase data

	β	se	robust se	p
Marginal				
rx1	-0.365	0.13	0.13	0.005
rx2	-0.121	0.22	0.22	0.590
rx3	-0.729	0.35	0.44	0.096
Conditional				
rx1	-0.365	0.13	0.13	0.005
rx2	0.208	0.22	0.21	0.310
rx3	-0.259	0.36	0.34	0.450

Table 11: Stratum specific fits to the dnase data

not, the important change is to model a single treatment effect for events 3–5. For the conditional model, the change is effected by capping the strata variable `enum` at a value of 3.

Table 10 shows the result of simple fits to the rhDNase data, and table 11 the results for more complicated models. The most acceptable model is the marginal model with 3 covariates. In this we see an apparent lessing of the treatment effect in stratum 2 and an increase in stratum 3. The individual contrasts between rx1/rx2 and rx1/rx3 are not significant, however, with $p = .18$ and .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. No evidence for non proportional hazards was observed.

The results of analyzing the recurrent events from the double-blind trial suggested a possible diminishing treatment effect; however no conclusion could be made as a result of too little information. However, long term effect of rhDNase was estimable from data collected during a post double-blind observation period. At the end of the 169 day trial, the treatment was determined to be 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. Standard errors are again based on the robust estimate $\widetilde{D}'\widetilde{D}$. Since the patients in the double-blind trial were enrolled during a six month pe-

Exacerbation Number	Relative Risk	Standard Error
1	0.72	0.10
2	0.78	0.18
> 2	0.54	0.38

Table 12: Estimated risk of exacerbation: double-blind and follow-up periods.

	Double-blind		Follow-up
	Placebo	rhDNase	rhDNase
Mean FVC (percent)	78	78	78
Mean FEV1 (percent)	61	61	61
Mean Age	18	19	19

Table 13: Characteristics of Patients: Double-Blind and Follow-up periods.

riod beginning in February 1992, data from months 7 through 12 of the follow-up period were used in the analysis in order 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 through 12 of the follow-up period. The results of fitting a marginal model to the time until each exacerbation are summarized in Table 12. The relative risk estimates of 1st 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% 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, age and lung function at baseline were compared to the values at month 7 of the follow-up, the results are displayed in table 13. No difference in patient characteristics was seen.

Based on this analysis, rhDNase produced a significant and sustained reduction in the risk of pulmonary exacerbations in patients with cystic fibrosis.

References

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **10**, 1100–20.

[2] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1992) *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

[3] Barlow, W.E. and Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika* **75**, 65–74.

[4] Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–47.

[5] Cain, K.C. and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* **40**, 493–99.

[6] Chappell, R. (1992). A note on linear rank tests and Gill and Schumacher's tests of proportionality. *Biometrika* **79**, 199–201.

[7] Chen, C.H. and Wang, P.C. (1991). Diagnostic plots in Cox's regression model. *Biometrics* **47**, 841–50.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion). *J. Royal Stat. Soc. B* **34**, 187–220.

[9] Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis.* Wadsworth, Belmont.

[10] Cochran, W.G. (1976). *Sampling Techniques, third edition.* Wiley, New York.

[11] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Am. Stat. Assoc.*, **72**, 27–36.

[12] Dickson, E.R., Grambsch, P.M., Fleming, T.R, Fisher, L.D. and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology* **10**, 1–7.

[13] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

[14] Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

[15] Gill, R. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.

[16] Grambsch, P.M. and Therneau, T.M. (1994).multi3 Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–26.

[17] Grambsch, P., Therneau, T.M., and Fleming T.R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51**, 1469-1482.

[18] Harrell, F. (1986). The PHGLM procedure. *SAS Supplemental Library User's Guide* Version 5. Cary, NC: SAS Institute Inc.

[19] Harrell, F.E., Pollock, B. G. and Lee, K. L. (1987). Graphical methods for the analysis of survival data. In *Proceedings of the Twelfth Annual SAS Users Group International Conference*, p 1107-1115. SAS Institute, Inc., Cary, NC.

[20] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[21] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221–33.

[22] Kalbfleisch, J.D. and Prentice R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

[23] Laird, N.M. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Am. Stat. Assoc.* **76**, 231–40.

[24] Lee, E.W., Wei, L.J. and Amato D. (1992). *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, 237–247, Kluwer Academic Publishers, Netherlands.

[25] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-30.

[26] Lindor, K.D., Dickson, E.R., Baldus, W.P., Jorgensen, R.A., Ludwig, J., Murtaugh, P.A., Harrison, J.M., Wiesner, R.H., Anderson, M.L., Lange, S.M., LeSage, G., Rossi, S.S and Hofman, A.F. (1994). Ursodeoxycholic acid in the treatment of primary biliary cirrhosis. *Gastroenterology* **106**, 1284–90.

[27] Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1990) Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Comm Stat Theory Meth.* **19**, 821–45.

[28] Lin, D.Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J. Am. Stat. Assoc.* **86**, 725–728.

[29] Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Stat. Assoc.* **84**, 1074–78.

[30] Lin, D.Y. and Wei, L.J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* **1**, 1–17.

[31] Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine* **13**:2233-47.

[32] Lin, D.Y., Wei, L.J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–72.

[33] Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics* **51**, 524-32.

[34] McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models.* Chapman and Hall.

[35] Miller, R.G. (1981), *Survival Analysis*, Wiley, New York.

[36] Moreau, T., O'Quigley, J. and Mesbah, M. (1985). A global goodness-of-fit statistic for the proportional hazards model. *Applied Statistics* **34**, 212–218.

[37] Moss, A.J. and the Multicenter Diltiazem Trial Research Group (1988). The effect of diltiazem on mortality and reinfarction after myocardial infarction. *New England J. of Medicine*, **319**, 385–92.

[38] Nagelkerke, N. J. D., Oosting, J., and Hart, A. A. M. (1984). A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics* **40**, 483–486.

[39] Oakes, D. (1992). *Frailty models for multiple event times.* In Klein, J.P and Goel, P.K. (eds), Survival Analysis, State of the Art, 415. Kluwer Academic Publishers, Netherlands.

[40] O'Quigley, J. and Pessione, F. (1989). Score tests for homogeneity of regression effects in the proportional hazards model. *Biometrics* **45**, 135–144.

[41] Prentice, R.L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika*, **79**, 495–512.

[42] Prentice, R.L. and Cai, J. (1991). *Marginal and conditional models for the analysis of multivariate failure time data.* In Klein, J.P and Goel, P.K. (eds), Survival Analysis, State of the Art, 393–406. Kluwer Academic Publishers, Netherlands.

[43] Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373–79.

[44] Reid, N. and Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72**: 1–9.

[45] Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–53.

[46] Smith, P.J. and Hietjan, D.F. (to appear). Testing and adjusting for overdispersion in generalized linear models. *J Royal Statistical Soc, Series C.*

[47] Therneau, T.M., Grambsch P.M. and Fleming, T.R. (1990). Martingale based residuals for survival models. *Biometrika* **77**, 147–60.

[48] Uitti, R.J., Ahlskog, J.E., Maraganore, D.M., Muenter, M.D., Atkinson, E.J., Cha, R.H. and O'Brien, P.C. (1993). Levodopa therapy and survival in idiopathic Parkinson's disease: Olmsted County Project. *Neurology* **43**, 1918–26.

[49] Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Stat. Assoc.* **84**, 1065–73.

[50] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–38.

[51] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika* **50** 1–26.