EVALUATION OF METHODS FOR ANALYZING MULTIPLE TIME-TO-FAILURE DATA

by
Katherine Harris Lipschutz
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2138T

December 1994

EVALUATION OF METHODS FOR ANALYZING MULTIPLE TIME-TO-FAILURE DATA

by

Katherine Harris Lipschutz

A dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill

1994

Approved by:

Advisor

Reader

Reader

Reader

Reader

Katherine Harris Lipschutz. Evaluation of Methods for Analyzing Multiple Time-to-Failure Data (Under the Direction of C. E. Davis).

ABSTRACT

In many clinical trials the event of interest is nonfatal and can recur. The standard analysis models the time to the first occurrence of the endpoint; however, this analysis ignores recurrences which might provide additional information and enhance power. Six models for analyzing time-to-event data where events are nonfatal and can recur are evaluated using simulations and examples. The models differ in their adjustments for dependency among endpoints, in their definition of event time and in their risk sets. Three of the models make implicit assumptions about the structure of the within subject correlations and two of these actually introduce correlation between event times within subjects by modelling total time since study start instead of gap time between events (Andersen and Gill 1982; Prentice, Williams and Peterson 1981). The true significance levels of tests based on these models are shown to exceed nominal levels in many cases. A fourth model estimates the correlation of the observations without specifying the structure and this model preforms well (Wei, Lin and Weissfeld 1989). The size of the test remains near nominal levels in most situations; however, this method models total time since

study start and assumes that all patients are at risk for all events as long as they remain in the study. These features make interpretation of model parameters questionable. Another recent method (Pepe and Cai 1993) adjusts for the correlation within subjects without specifying the structure and has the advantage of restricting the risk set so that patients are at risk for event k only if they have experienced k-1 events. The final method evaluated is a modification of the method proposed by Pepe and Cai that models gap times between events rather than total event times. These last two methods are extremely sensitive to small sample sizes and heavy censoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AG	Andersen and Gill Method
CHF	Congestive Heart Failure
CONSENSUS	Cooperative North Scandinavian Enalapril Survival Study
GLM	General Linear Model
GLMM	General Linear Multivariate Model
GVHD	Graft Versus Host Disease
MI	Myocardial Infarction
MLE	Maximum Likelihood Estimate
NYHA	New York Heart Association
PC	Pepe and Cai Method
PC_m	Modified Pepe and Cai Method
PWP	Prentice, Williams and Peterson
PWP_2	Prentice, Williams and Peterson Model 2
PWP_3	Prentice, Williams and Peterson Model 3
RR	Relative Risk
SOLVD	Study of Left Ventricular Dysfunction
WLW	Wei, Lin and Weissfeld

CHAPTER I
INTRODUCTION

Survival analysis is used in clinical trials to compare two or more treatments with respect to differences in survival distributions. An important feature of survival type problems is that some subjects may be censored; *i.e.*, the event of interest does not occur during the study. This has led to the development of special statistical techniques that make use of the partial information contributed by censored observations.

Typically survival analysis is concerned with the time to occurrence of a **single** event. However, there has been increased interest recently in methods for analyzing time-to-event data when multiple failures can occur during follow-up of an individual patient. Multiple failures within a subject can either be occurrences of events of different types or recurrences of a single event type. Although methods for both forms of multiple failures are reviewed here, the primary focus of this paper is on recurrent events of a single type. Some examples of recurrent events include hospital admission for congestive heart failure (CHF) (SOLVD 1991), myocardial infarction in patients with low ejection fraction (SOLVD 1992), detection of tumors in bladder cancer patients (Wei, Lin and Weissfeld 1989) and occurrence of infections in bone marrow transplant patients (Pepe and Cai 1993). In clinical studies of recurrent events it might be of interest to evaluate treatment or other covariate effects on all events, rather than just the first event. Several methods have been proposed to analyze these types of data and, of these, six have been chosen for evaluation in this report.

It seems logical that second and subsequent events contribute new information that could improve the likelihood of detecting a treatment effect. This increase in power will be examined and situations will be identified in which the recurrent endpoint methods actually have less power than an analysis of the time to the first event only.

As with the analysis of repeated observations in any longitudinal study setting, the dependency of observations within individuals must be addressed. Along these lines, most existing methods for recurrent survival data can be classified into two broad categories: 1) methods that explicitly model the dependency within subjects and 2) methods that treat the dependency as a nuisance and adjust for it in the estimation of the covariance matrices. Most approaches that model the dependency of event times through a regression model rely on counting process theory (Andersen and Gill 1982; Prentice, Williams and Peterson 1981; Gail, Santner and Brown 1980). Other methods, such as frailty models, use random effects to model the dependency. Since the validity of inferences based on these models depends on whether the model assumptions are correct, it is preferable to use methods that treat the dependency as a nuisance and estimate the correlation without making assumptions about the structure. The performance of the methods will be compared under various dependency structures.

Another feature that varies among multiple event survival methods is the definition of the risk set for a recurrent event. Some methods specify that all subjects remaining in the study at time t are at risk for the k^{th} event, even subjects that haven't experienced the $(k-1)^{th}$ event. Other methods restrict the risk set for the k^{th} event to include only those patients who have experienced (k-1) events. A third type of risk set imposes the additional restriction that the subjects must have experienced (k-1) events *before* time t, the time of the k^{th} event. The effects of these risk set definitions on the interpretation of the

model parameters will be discussed and their effects on the performance of the methods will be evaluated.

The six methods will be evaluated on the basis of their actual significance levels and power. The main questions that will be addressed are: 1) What is the effect (size of the test) of incorrectly assuming a specific dependence structure? 2) How do tests using unrestricted risk sets perform? 3) Is there improvement in or detriment to power by combining information over all events relative to a time to first event analysis? 4)Since all six models rely on asymptotic normality for inference about regression parameters what are the small sample properties of the models? In addition, the methods will be illustrated using data from three cardiovascular clinical trials. Finally, the results will be summarized and recommendations concerning the choice of an appropriate method will be made.

CHAPTER II
SURVIVAL ANALYSIS APPROACH TO MULTIPLE FAILURE DATA

This section begins with a broad overview of the field of survival analysis, with particular emphasis on semi-parametric methods that have been developed to examine the effects of covariates on time-to-event data. The problem of multiplicity in clinical trials is introduced, followed by a review of methods designed specifically for analyzing recurrent event time-to-failure data. Five methods are reviewed in depth and a new method is proposed.

A. Survival Analysis Approach to Time-to-Event Data, Single Outcome

The primary purpose of survival analysis in clinical research is to assess the relationship of a set of covariates, such as treatment regimen, with a positive valued random variable, usually time to the occurrence of a specified event. Examples of such events in the clinical research area include hospital admission for congestive heart failure and myocardial infarction in patients with left ventricular dysfunction (SOLVD 1991; SOLVD 1992), tumor occurrence in patients with bladder cancer (Wei, Lin and Weissfeld 1989) and post-transplant infections (Pepe and Cai 1993). In clinical trials, the event may not occur in some subjects during the study period; in these cases the event is said to be censored. Censoring is an important characteristic of time-to-event data that has necessitated the development of specialized statistical methods. Standard parametric or nonparametric methods cannot be routinely applied because they cannot make use of the partial information contributed by a censored patient. Even in the absence of censoring (i.e., when all subjects experience the event) survival analysis methods are often preferred over standard methods for

continuous data because of the nature of time-to-event random variables: They are always positive valued and their distribution is usually skewed (Lee 1980, p.2). Similarly, the use of methods appropriate for binary outcomes, such as logistic regression, would result in the loss of valuable information. In fact, in the absence of censoring this analysis becomes totally non-informative.

i. Censoring

If an event is not observed for a patient by the end of the study, that patient's exact event time is not known and the patient's outcome is considered right-censored. There are three types of right-censoring: Type I, Type II and random censoring. Type I censoring occurs when patients are started together and observed for a prespecified duration. In Type II censoring patients are started together and observed until a prespecified number of events have occurred (Miller 1981, pp. 3-4). Both of the above types of censoring are considered fixed, in that all patients are observed for the same length of time. These types are also known as singly censored since all censored subjects have the same censoring time (Lee 1980, pp. 2-3). Data terminated by Type I censoring is also known as truncated data, while data subjected to Type II censoring is simply referred to as censored data (Elandt-Johnson and Johnson 1980, p. 50).

The feature that distinguishes random censoring from fixed censoring is that patients enter the study at different times. In this case, if the study is ended at a predetermined date then it is likely that many patients will have different censoring times, even if they complete the study. Other sources of random censoring, also known as progressive time censoring, include patients who are withdrawn or lost from the study and were

event free at the last observation (Lee 1980, p. 3). Since random or progressive censoring is the most common type of censoring in clinical trials, it will be the type used in the following applications. Furthermore no distinction will be made between planned censoring (end of study) and unplanned censoring (withdrawn or lost to follow-up).

The random censorship model can be defined as follows. For subject $i=(1,..n)$, assume that T_i (event time) and U_i (censoring time) are each independent identically distributed (i.i.d.) random variables; and

$$X_i = \min(T_i, U_i)$$
$$\delta_i = I_{(X_i = T_i)}$$

We observe (X_i, δ_i) where X_i are i.i.d. random variables with some known or unknown density function and δ_i contain the censoring information (Fleming and Harrington 1991, p. 90).

Finally, for the methods used here, it is necessary to assume that censoring times are independent of the event times in the sense that the hazard rate of the time-to-event variable, T, is the same as the hazard rate of T in the presence of censoring (Fleming and Harrington 1991, p. 20):

$$P\{s \le T < s+ds \mid T \ge s\} = P\{s \le T < s+ds \mid T \ge s, U \ge s\}.$$

This assumption seems reasonable when censoring is the result of random study entry or losses to follow-up but can be violated when study withdrawal is related to treatment or outcome. For example, consider a survival trial comparing an active agent with placebo in the treatment of congestive heart failure. If patients are withdrawn from the

trial and censored due to serious adverse experiences, which are actually symptoms of heart disease (implying worsening condition), then survival may be overestimated for these patients. This situation would be especially troublesome if nonindependent censoring occurred at different rates in the two treatment groups. Adverse experiences due to lack of efficacy resulting in withdrawal may occur more often in the placebo group of a study with an effective experimental treatment. On the other hand, adverse experiences resulting from toxicity to the active treatment may also cause patients to withdraw. When censoring time is not independent of event time, most common statistical approaches to survival analysis result in biased estimates and inaccurate inference. In some situations an intention-to-treat approach to the analysis of survival data can be used in an attempt to remove bias in assessing efficacy. This approach attributes all events to the treatment group to which the patient was assigned even if the subject is later removed from the treatment. Obviously, the intention-to-treat approach can only be used when the event time is known and therefore it does not completely remove the bias due to nonindependent censoring.

ii. Functions Describing Survival Data

Let f(t) be the probability density function used as the distributional model for the random variable T, time to an event. There are two additional important functions used to characterize survival data: the survival function S(t) and the hazard function $\lambda(t)$. The survival function is defined as:

$$S(t) = Prob(T > t) = 1 - F(t).$$

This is the probability that the event occurs after time t, in other words the probability a patient will be event free at least until time t. S(t) is also known as the cumulative survival rate.

The hazard function, $\lambda(t)$, is

$$\lambda(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

and can be interpreted as the probability that the event occurs within a small interval following t, given the event has not occurred before time t. The hazard function can also be interpreted as the instantaneous relative failure rate at time t (Elandt-Johnson and Johnson 1980, p. 51). The area under $\lambda(t)$ is defined as the cumulative hazard function and can be expressed in terms of the survival distribution:

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{1-F(u)}du = -\log[1-F(u)]\Big|_0^t$$
$$= -\log[1-F(t)] = -\log S(t)$$

Similarly,

$$S(t) = e^{\int_0^t \lambda(u)du}$$
$$= e^{-\Lambda(t)}$$

The probability density function in survival analysis is interpreted as the unconditional probability of an event within an arbitrarily small interval around time t and can be expressed as a function of the survival function

and the hazard rate function (Elandt-Johnson and Johnson 1980, p. 51):

$$f(t) = \lambda(t) S(t).$$

As long as T has a density, (i.e., the distribution function is continuous) S(t), λ(t) and f(t) are three ways of describing the same distribution, and if one is known the other two functions can be derived (Lee 1980, p. 9).

iii. Maximum Likelihood for Parametric Models

There are many parametric methods used for the estimation and analysis of censored event-time data. These methods specify the functional form of the survival, hazard or cumulative hazard function. Maximum likelihood methods can then be used to estimate the unknown parameters of the functions. In the survival analysis setting, the partial information contributed by censored observations must be incorporated into the maximum likelihood functions.

Assume random censoring, although slight modification to the likelihoods can incorporate Type I or Type II censoring as well (Miller 1981, p. 17). If X_i represents an observed failure time for subject i then the i^{th} contribution to the likelihood is the density function for T evaluated at X_i. If X_i represents a censored time, the contribution to the likelihood is the probability that $T_i > U_i$ or equivalently that $T_i > X_i$. Therefore, the joint likelihood over all $i = 1,2,\ldots,n$ is:

$$L = \prod_{i:\delta_i=1}^{n} f(X_i) \prod_{i:\delta_i=0}^{n} S(X_i).$$

Remembering that $f(t) = \lambda(t)S(t)$ and $\Lambda(t) = -\log S(t)$, the log likelihood is given by:

$$\log L = \sum_{i:\delta_i=1}^{n} \log\lambda(X_i) - \sum_{i=1}^{n} \Lambda(X_i). \qquad (1)$$

The distribution of the censoring times is not included in the above likelihood. Since the censoring distribution is assumed to be independent of the survival time distribution, it contains no information about the unknown survival function parameters. The censoring time component of the likelihood would be multiplicative and treated like a constant when maximizing L (Miller 1981, p. 17). Thus, the estimates of survival distribution parameters can be obtained from the likelihood without specifying the censoring distribution.

iv. Common Parametric Models

Often, the true functional form of the hazard distribution is unknown, and models must be selected to characterize the distribution. Models describing event or failure time data must have hazard functions which are appropriate for the application. For example, suppose it is known that for large t (*i.e.*, long term follow-up) the hazard rate probably doesn't decrease over time, while for small t the hazard might be expected to be

high initially and decrease (*e.g.*, to represent an initial high risk period following surgery).

The exponential distribution is a single parameter distribution that assumes a constant hazard rate and no memory; thus, failure is a random event independent of time (Lee 1980, p.159). This distribution describes a poisson process. The probability density function, survivorship function and hazard function are given below:

$$f(t)=\lambda e^{-\lambda t}, \qquad t\geq 0,\ \lambda>0,$$

$$S(t)=e^{-\lambda t}, \qquad t\geq 0,$$

$$\lambda(t)=\lambda, \qquad t\geq 0.$$

Because of the constant hazard the exponential distribution is not appropriate for survival situations that involve aging or wearing out. Clinical situations that may require non-constant hazard rates include cancer trials where the hazard may increase over time, and post-surgical mortality studies where hazard rates might be high initially and then decrease. Although restrictive in applications, the exponential distribution is important because it is the basis for several more general distribution functions used to model survival (Lee 1980, p.158).

The Weibull distribution is a generalization of the exponential distribution that does not assume a constant hazard rate over time. The Weibull distribution is characterized by two parameters, α and λ, representing shape and scale, respectively. The probability density function, survivorship function and hazard function are given below:

$$f(t)=\lambda\alpha(\lambda t)^{\alpha-1}\exp[-(\lambda t)^{\alpha}], \quad t\geq 0,\ \alpha,\lambda>0,$$

$$S(t)=\exp[-(\lambda t)^{\alpha}],$$

$$\lambda(t)=\lambda\alpha(\lambda t)^{\alpha-1}.$$

The Weibull distribution has a hazard rate that is increasing for values of $\alpha > 1$ (positive aging), decreasing for $\alpha < 1$ (negative aging) and constant for $\alpha = 1$ (exponential model) (Lee 1980, p. 168). Figures 1-3 illustrate the hazard, density and survival functions, respectively, for the Weibull distribution with $\alpha=0.5, 1.0, 2$ and $\lambda=(2)^{\frac{1}{2}}, 1, \pi/4$. Note that all of these distributions have an expected value of one.

Figure 1

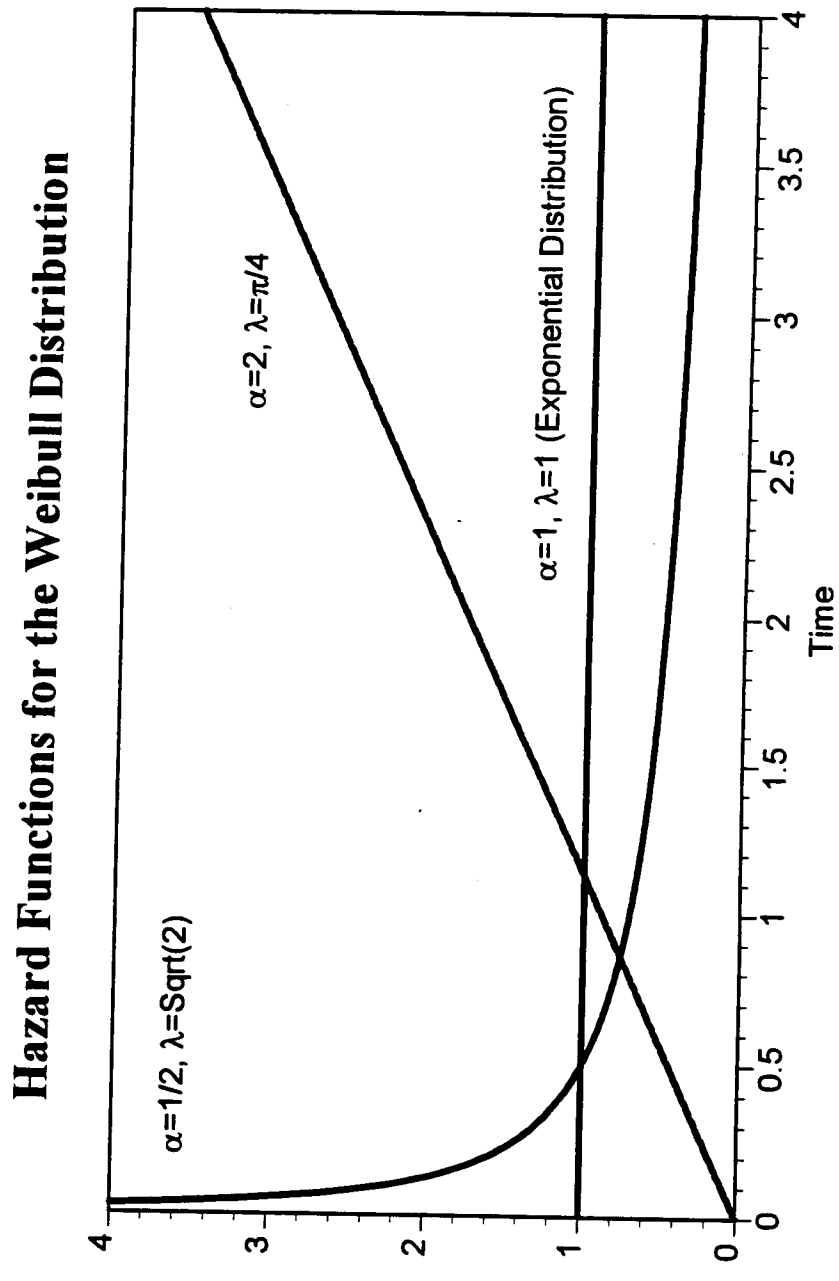

Hazard Functions for the Weibull Distribution
α=1/2, λ=Sqrt(2)
α=2, λ=π/4
α=1, λ=1 (Exponential Distribution)
Time
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4

Figure 2

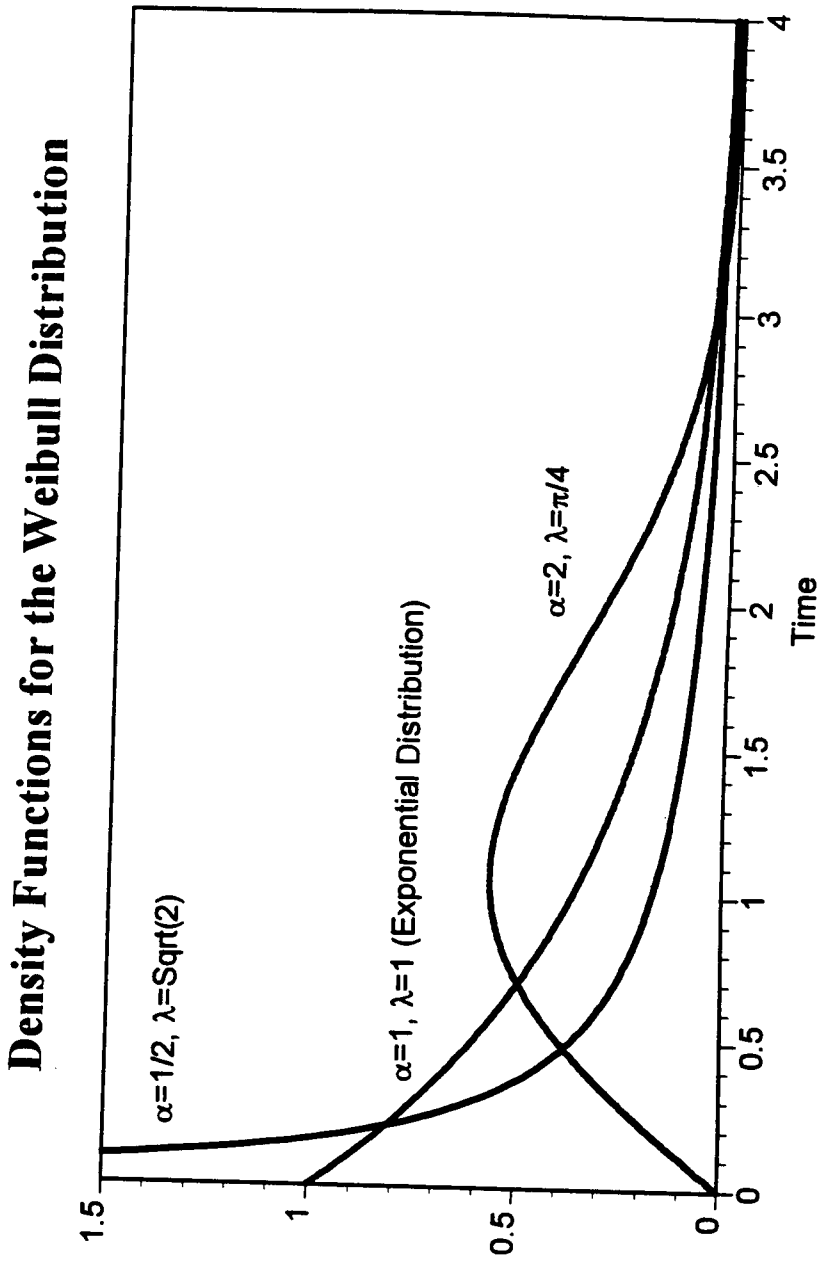
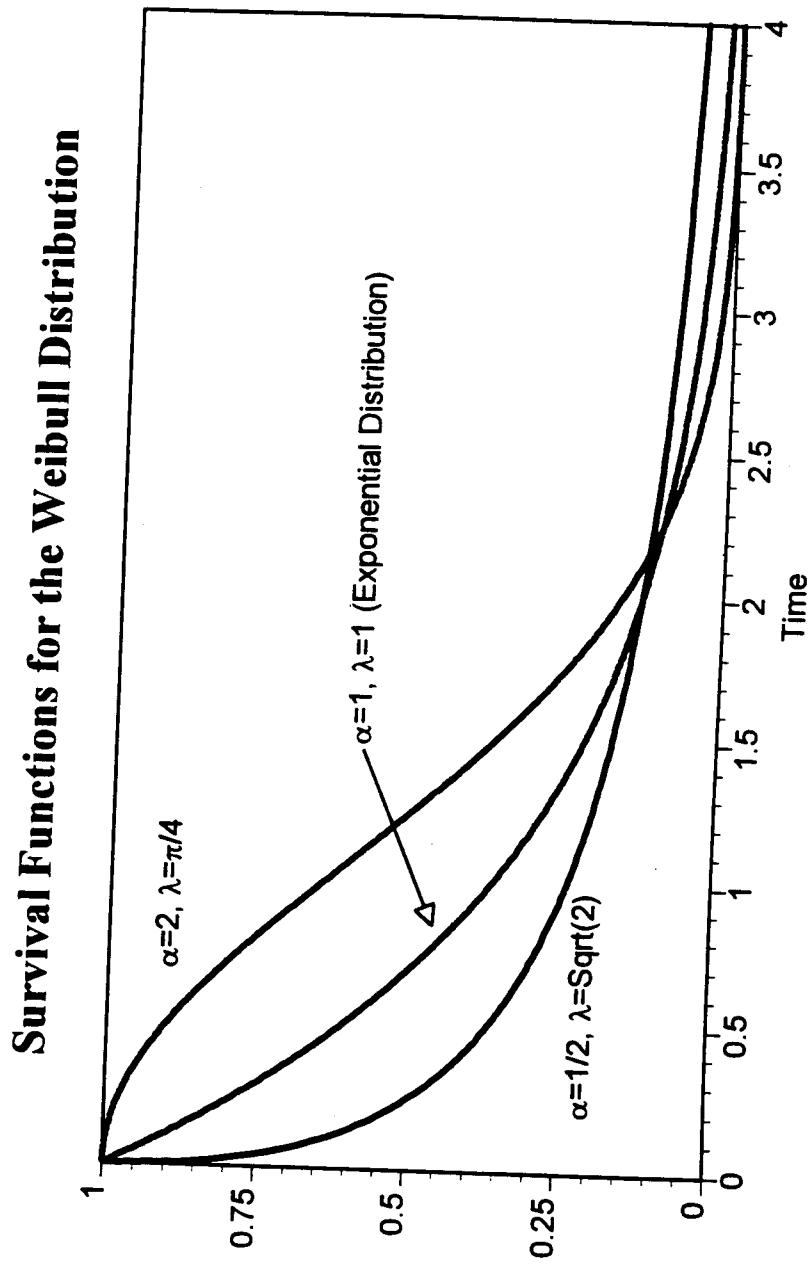

Density Functions for the Weibull Distribution
α=1/2, λ=Sqrt(2)
α=1, λ=1 (Exponential Distribution)
α=2, λ=π/4
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time

Figure 3


Survival Functions for the Weibull Distribution
α=2, λ=π/4
α=1, λ=1 (Exponential Distribution)
α=1/2, λ=Sqrt(2)
1
0.75
0.5
0.25
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time

Other parametric distributions are also sometimes used to describe survival data. The lognormal distribution is used for survival patterns with initially increasing and then decreasing hazard rate. The gamma distribution is used less in clinical research and is applicable for survival patterns where the hazard rate is monotonically decreasing from infinity ($\gamma < 1$) or increasing from zero ($\gamma > 1$) to a constant value, λ, as time increases (Lee 1980, p. 169). The linear exponential and Gompertz distributions describe survival patterns with initial constant hazard that increases linearly with time in the case of the linear exponential, and exponentially with time for the Gompertz (Lee 1980, p.179).

For known survival models, standard maximum likelihood techniques described in section II.A.iii can be used to obtain parameter estimates and confidence intervals; however, for most distributions the maximum likelihood equations are complicated when there are censored data. The exponential distribution is one of the few distributions with a closed form solution for its parameters in the presence of censoring (McCullagh and Nelder 1989, p. 423).

v. Regression for Parametric Models

For any of the parametric models, heterogeneity within the sample can be explained by adding predictor variables in a regression setting. Let $\mathbf{Z}' = (z_1, z_2, \ldots, z_s)$ represent a vector of concomitant variables, $\hat{\boldsymbol{\beta}}$ the associated parameter estimates and $\lambda(t)$ the reference or baseline hazard rate (that is, the hazard rate for a hypothetical individual with $\mathbf{z} = \mathbf{0}$).

Most regression models for survival data separate the time and covariate effects, and also assume proportional hazards (Lee 1980, p. 319). In this

case the model can be defined as:

$$\lambda(t;\mathbf{z};\boldsymbol{\beta})=\lambda(t)g(\mathbf{z};\boldsymbol{\beta})$$

This model assumes that the ratio of the hazards for two individuals is constant over time. A popular approach to assessing the effects of the covariates on the hazard or survival function is to allow the hazard function to be multiplied by $\exp\{\boldsymbol{\beta}'\mathbf{z}\}$. The assumptions required for this model are: 1) The form of the underlying hazard function is correctly specified; 2) The relationship between predictors and log hazard is linear; and 3) The relationship between predictors and response is the same at each t; *i.e.*, proportional hazard. Therefore, predictor by time interaction indicates nonproportional hazard.

To summarize, the model implies that the ratio of hazards for two individuals depends on the difference between their linear predictors and is a constant independent of time as long as there are no time-dependent covariates (McCullagh and Nelder 1989, p. 421). The model is specified as:

$$\lambda(t;z)=\lambda(t)\exp(\boldsymbol{\beta}'\mathbf{z})$$

The parameters λ and $\boldsymbol{\beta}$ can be estimated by maximum likelihood using a formula similar to equation (1) found in section II.A.iii, with the addition of $\exp(\boldsymbol{\beta}'\mathbf{z})$. This model is quite popular because no restrictions need to be placed on the value of $\boldsymbol{\beta}$, since $\exp(\boldsymbol{\beta}'\mathbf{z})>0$, for any $\boldsymbol{\beta}$ and z. In contrast, for additive or linear-exponential models the values of $\boldsymbol{\beta}$ must be restricted to ensure $\boldsymbol{\beta}'\mathbf{z} \geq 0$ since by definition the hazard function can never be negative (Elandt-Johnson and Johnson 1980, pp. 353-

354). The parametric proportional hazards regression model can be generalized to include polynomial effects (*e.g.*, log-quadratic) as well as time-dependent covariates. Interaction terms can allow differential treatment effects for different covariates.

Throughout most of this paper it is assumed that the form of the survival distribution is not known and therefore subsequent emphasis will be given to nonparametric approaches to estimating and testing survival models. In the regression setting, leaving the functional form of $\lambda(t)$ unspecified and treating it as a blocking factor defined only at points where deaths occur results in Cox's model (McCullagh and Nelder 1990, p. 421). Cox's model is described in section II.A.vii.

vi. Nonparametric Estimation of Survival Function

Nonparametric methods are more efficient than parametric methods for estimating survival functions when the distribution of the survival times is not known (Lee 1980, p. 75). The classical method for estimating the survival function, S(t), using the partial information from censored observations and without assuming a parametric form is the actuarial method (Miller 1981, p.39). This method is used when events are grouped within time intervals; *i.e.*, when exact event times are not known or are not used. Estimates obtained with this method are also known as life-table estimates. A detailed discussion of this method can be found in Elandt-Johnson and Johnson (1980), pp 154-172.

The Kaplan-Meier product-limit (PL) estimate is a nonparametric estimator of the survival function when the exact event and censoring times are known. The product limit estimator is similar to the actuarial

estimator in that they are both formed by multiplying together a sequence of estimates of conditional probabilities of survival through a series of intervals (Kaplan and Meier, 1958). However, in contrast to the actuarial estimator, the PL estimator uses the individual ordered event or censoring times; Hence, the product limit estimator uses all the available information. Furthermore, the PL estimator does not require the actuarial estimator assumption that censored subjects are, on average, exposed to risk for half of an arbitrary time interval (Elandt-Johnson and Johnson 1980), *i.e.*, that censoring is uniformly distributed in the interval (Lee 1980 p. 91). In practice, intervals used in the product limit estimate are defined by event times only, since the survival estimates are unchanged at times of censoring only.

The PL estimator of survival probability through event time i can be summarized as the probability of surviving through event time i-1 multiplied by the observed survival rate for event time i (Lee 1980, p. 77). Let n be the total number of individuals with survival times, including censored and failure events, and order the n survival times such that

$$t_1 \leq t_2 \leq \ldots \leq t_n.$$

The PL estimator is then given by

$$\hat{S}(t) = \prod_{t_i \leq t} \left[\frac{n-i}{n-i+1} \right]^{\delta_i}.$$

So far we have assumed that the product-limit method uses the exact time of the event. By exact time we mean time measured along a scale that is fine enough (*e.g.*, an hour or a day) so that the probability that

two failure events occur at the same time is low. In the case of ties, ranks can be assigned arbitrarily to the tied observations and the cumulative survival for that time is the smallest of the $S(t_i)$ (Lee 1980, p. 76). The PL method assumes that a patient censored on a certain day survived until the end of that day, which in turn implies that if an event and censor time are both recorded on day i, the censored individual is included in the risk set for day i.

The cumulative hazard estimator corresponding to the PL estimator,

$$\hat{\Lambda}_2(t)=\sum_{t_i\le t}\left[\frac{1}{n-i+1}\right]^{\delta_i}$$

is similar to the Nelson's cumulative hazard estimator,

$$\hat{\Lambda}_1(t)=\sum_{t_i\le t}\left[-\log\left(1-\frac{1}{n-i+1}\right)\right]^{\delta_i}$$

because when x is small $\log(1-x)\approx -x$ (Miller 1981, p. 66).

It has been shown that both the Kaplan and Meier product-limit estimator and the Nelson estimator can be obtained using generalized or nonparametric maximum likelihood techniques. In addition both are consistent estimators (Fleming and Harrington 1991, p.121). Fleming and Harrington showed that the Nelson estimator has slightly smaller mean square error than the product-limit estimator (Fleming and Harrington 1984).

vii. Semi-parametric Regression Model

Cox introduced the semi-parametric proportional hazard regression model and later developed the associated partial likelihood approach to parameter estimation and inference (Cox 1975). These methods form the basis for many of the methods used to analyze survival data with recurrent failure events.

Cox's proportional hazard regression model is used to assess the relationship between a set of covariates and the hazard or survival function without specifying the functional form of the baseline hazard function. Cox's model is often referred to as semi-parametric. It is nonparametric in the sense that the form of the baseline hazard function is not specified but parametric with respect to specifying the effect of the covariates on the hazard function. The regressors are assumed to be linearly related to the log hazard and additive. This type of regression assumption must also be made for models that make additional parametric assumptions regarding the underlying hazard. This amount of parametric modelling is unavoidable when the objective of the analysis is to relate regressors to response.

Cox's model is popular because of it's robustness. For estimating and testing regression coefficients, the Cox model is nearly as efficient as parametric models even when all the assumptions about the parametric model are true. When the parametric assumptions are not true, the Cox analysis is more efficient than the parametric analysis (Harrell 1992).

The model can be defined as follows: Let $j=(1,..s)$ represent the s covariates and $i=(1,..n)$ indicate subject. Define T_i (event time) and U_i (censoring time) and (X_i, δ_i) as in section II.A.i. As in the parametric

proportional hazards regression model let $z' = (z_1, z_2, \ldots, z_s)$ represent a 1xs vector of concomitant variables, $\hat{\beta}$ the associated parameter estimates and $\lambda_0(t)$ the reference hazard rate, or the hazard rate when all z's equal zero. Representing the relationship between the survival time distribution and the covariates, the hazard for the i^{th} subject is as follows:

$$\lambda_i(t;z) = \lambda_0(t)\exp(\sum_{j=1}^{s} \beta_j z_{ij}).$$

The hazard rate is then the product of a scaler and the baseline hazard rate, $\lambda_0(t)$, where the scaler depends on the values of the regression coefficients and covariates (Miller 1981, p. 120). The relative hazard ratio for any two values of a covariate, $\exp(\beta' z)$, is constant over time; *i.e.*, the hazard functions are proportional. This in turn results in survival functions that are powers of one another (Lee 1980, p. 307).

An important characteristic of Cox's proportional hazards model is that $\lambda_0(t)$ is arbitrary, which implies that standard maximum likelihood techniques cannot be used to estimate the regression coefficients. To obtain estimates of and inference on the regression parameters, Cox (1975) constructed the following partial likelihood based on conditional probability arguments. He originally assumed continuous survival times. Let i index the n subjects as before, and define R_i as the set of individuals at risk of failing at time t_i. The partial likelihood is then given by

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta' z_i}}{\sum_{l \in R_i} e^{\beta' z_l}} \right]^{\delta_i}.$$

Estimates of β are found by applying maximum likelihood techniques to the partial likelihood and then using the Newton-Raphson method of iteration. Inference on the β is obtained from the score vector and sample information matrix or from likelihood ratio tests. In general, the asymptotic normality of the score vector from maximum likelihood is derived from the central limit theorem for independent observations; however, the score vector obtained from the partial likelihood is not the sum of independent observations (Fleming and Harrington 1991, p. 148). Tsiatis (1981) proved that $\hat{\beta}$ is asymptotically normal.

Application of partial likelihood from Cox's model requires the assumption that censoring times and covariates of censored patients provide no information with respect to β. If this is the case, censoring is said to be noninformative with respect to the regression parameters. Note that independence of censoring times does not guarantee that censoring is uninformative. T and U can both be i.i.d. random variables that depend on β (Fleming and Harrington 1991, p. 139).

Since the model assumes survival times are continuous, modifications must be made to handle tied failure times. Cox (1975) derived a likelihood for tied event times but his approach was computationally difficult. For this reason Breslow (1975) suggested a simpler approximate partial likelihood for tied observations. Breslow's estimate assumes an underlying continuous survival function and is based on the joint likelihood of β and $\lambda_o(t)$ assuming all censored observations occurring in interval $[t_i, t_{i+1}]$ occur at t_i, and that the hazard is constant between $[t_i, t_{i+1}]$.

Kalbfleish and Prentice (1973) formalized Cox's proportional hazard regression model by deriving the same likelihood based on a marginal model of the ranked failure times (Miller 1981). They extended their derivations to incorporate tied observations by considering all possible rank arrangements of the survival times (Lee 1980, p. 318). Since this estimate may be difficult to compute, Breslow's approximation is often preferred if data are not too heavily tied.

Cox's model can also be expressed as a multiplicative intensity model using counting process notation. Instead of modelling the conditional hazard for T, the intensity process, I(s), is modelled for N(t) (Fleming and Harrington 1991). If $\{N(t) : t \geq 0\}$ is a counting process representing the number of events that have occurred by time t and $Y(s) = 1$ at time s if a failure can be observed and $Y(s) = 0$ otherwise, then

$$I(s) = \lambda_0 \exp(\beta' z) Y(s).$$

Fleming and Harrington provide proofs that the score vector from Cox's partial likelihood is asymptotically multivariate normal based on martingale theorems applied to the counting process approach to failure data. Counting processes are reviewed in detail in section II.C.ii.b.

The popular log rank test proposed by Peto and Peto (1972) for comparing two survival distributions is equivalent to the score test from the Cox model with one dichotomous covariate. The proportional hazards assumption necessary for the Cox model also applies to the log rank test. Additionally, the score test from the stratified Cox model, discussed below, is equivalent to the stratified log rank test (Harrell 1992).

One important feature of Cox's model is that it can adjust for covariates that are modelled through stratification. This is a valuable tool because the adjustment for stratification variables can be done through the Cox model without meeting the proportional hazards assumption, which implies that the baseline hazard differs across levels of the stratification variable. Stratification variables are usually polytomous. Let S = (1, . . k) represent a stratification variable with k levels; the stratified Cox model for the ith subject is

$$\lambda_i(t;z,s) = \lambda_s(t)\exp(\sum_{j=1}^{s} \beta_j z_{ij}).$$

In a stratification model, a common vector of model coefficients is fitted across strata resulting in a pooled estimate of the covariate effects. Separate log likelihood functions are fitted for each stratum, with each having the same beta vector. The likelihood functions are multiplied, or log likelihoods added over strata and the total log likelihood is maximized to obtain the stratified estimate of β (Harrell 1992). Multiplying the likelihoods in this way requires that the observations across strata be independent.

Another important feature of Cox's model is that it allows covariates in the model to be time dependent. Furthermore, time by covariate interaction terms can be used to check the proportional hazard assumption.

The stratified Cox model plays an important role in some of the multiple failure models discussed in section II.C.

viii. Estimation of Survival Function Based on Cox's Model

Cox's model can be used to obtain a nonparametric estimate of the survival function which is a generalization of Nelson's cumulative hazard estimator (Breslow 1972; Breslow 1974) and is obtained by maximizing the joint partial likelihood with respect to β and λ.

B. Multiplicity in Clinical Trials

Multiplicity in clinical trials is a very general concept that applies to many different situations. For example, multiplicity can refer to multiple tests of hypotheses in a group of related variables or sequential testing of the same hypothesis. This section will briefly discuss various types of multiplicity; the specific case of multiplicity in survival trials will be discussed in the next section.

i. Multiple Tests of Related Hypotheses

Performing multiple tests of related hypotheses within the same experiment raises the issue of controlling the experiment-wise alpha level. This type of multiplicity is often in the form of comparing two therapies with respect to a collection of related measurements or endpoints measured in the same sample of subjects. If each endpoint is analyzed with an individual univariate test, then it is important to address the issue of the experiment-wise error rate for the series of n related hypotheses within one experiment. One solution to this problem is to adjust the alpha levels, p-values or critical values for each individual test so that the overall experimentwise alpha level remains at a prespecified level. Ng et. al. (1993) provide a useful summary of these types of methods for analysis of multiple endpoints.

Another solution to the problem of multiple hypotheses is to combine the individual hypotheses into a single test by identifying one primary endpoint, forming a composite score consisting of several endpoints or analyzing a collection of related outcomes with a single test statistic using multivariate procedures. One of the drawbacks with these procedures is that they do not indicate which of the individual outcomes is contributing to a significant difference if one is observed. For this reason, a series of univariate tests against specific hypotheses is often desired.

ii. Multiple Treatment Groups

Multiplicity can also arise when there are more than two therapies administered within an experiment and there is interest in the pairwise comparison of all or a subset of treatments. Usually this situation requires adjustment of alpha levels for individual tests so that the alpha level for the overall comparison does not exceed a prespecified value.

iii. Multiple Looks at the Data

Multiplicity also refers to interim or sequential analysis where the same hypothesis is tested more than once as patients accrue over time. The important issue that arises in this situation is the effect of repeated analyses of data on the alpha level of the statistical test performed at the end of the study. Similar issues are related to repeated confidence interval and point estimation. Sequential designs have been developed where data are analyzed after each new observation or pair of observations. More practical group sequential methods have been developed where statistical tests are performed at prespecified intervals (Pocock 1977).

iv. Longitudinal Data Analysis : Multiple Outcomes per Subject and Other Correlated Response Situations

In longitudinal studies a series of observations is typically made on each subject and observations within the same subject are expected to be positively correlated. In this sense longitudinal data analysis may also be applied to groups or clusters of observations that are not necessarily observed over time, but may be positively correlated, such as individuals in a family or litter. In order to obtain correct statistical inference, the analysis must account for the correlations among a subject's (or group's) responses (Zeger and Liang 1986). For regression applications, point estimates of the parameters are unaffected by this correlation; however, the variances of the parameters, if estimated without taking the correlation into account, will be under-estimated.

When the outcomes in a longitudinal study are Guassian and time intervals across subjects are consistent, a general linear multivariate model (GLMM) can be used. These models will automatically cope with the correlations within subjects. In fact there is a large class of linear models based on the multivariate normal distribution available for the analysis of longitudinal data (Ware 1985). A repeated measurements design approach is a longitudinal data study in which the structure of the covariance matrix is assumed to have compound symmetry: all pairs of observations from one subject have the same correlation, observations are homoscedastic (*i.e.*, population variance is the same for all observations), and the variance of observations and covariance between pairs of observations are the same for all subjects. By assuming compound symmetry, the multivariate problem of longitudinal data (GLMM) can be reduced to a univariate analysis (GLUM).

For discrete and other non-gaussian outcomes there are few multivariate distributions and thus few likelihood-based methods for analysis. Liang and Zeger (1986) proposed a method to obtain model parameter estimates and variances using general estimating score-like equations (GEE) based on quasi-likelihood assumptions. The GEE method for longitudinal data was developed further by Zeger and Liang (1986). The approach can be summarized as follows. Since score equations are functions only of the expected value and variance of the response, we need only know these two properties and not the full distribution of responses to obtain them. Thus the quasi-likelihood approach specifies only the relationships (1) between the expected value of the outcome and the covariates, and (2) between the mean and variance. Parameter estimates obtained as solutions to the score-like equations have been shown to be asymptotically normal (McCullagh 1983).

In generating estimating equations for longitudinal data the structure of the covariance matrix for each subject must also be specified. A variance-covariance estimator is provided that is robust for the particular choice of the correlation structure and the estimator is consistent whether or not the working correlation matrix is correct. Models can be specified as marginal, conditional or mixed. For mixed models individual effects are usually represented as random effects. Conditional and mixed models both dictate the structure of correlation matrix for within subject responses. In addition, random effects models require distributional assumptions for the random effect. A correlation structure that is completely arbitrary can be used with the general estimating equation only if patients follow the same visit schedule and there are the same number of subjects per visit (Zeger and Liang 1986). The original GEE approach to the analysis of longitudinal data was not applicable to survival data because it did not account for censored observations.

However, several of the methods discussed in the following section adapted this approach to censored, multiple failure survival data.

C. Multiple Outcomes in the Survival Analysis Setting

There are situations in survival analysis where subjects may experience more than one failure. In these situations, it may be of interest to compare two or more treatment groups with respect to total failure experience and have the ability to control for additional covariates that may also affect the risk of failure. Using all available information from a study by accounting for all occurrences of the event or events rather than just a single occurrence may provide an increase in the power to detect treatment differences. If the treatment effect is consistent over all event occurrences it may also be desirable to provide a meaningful estimate of the average treatment effect over all occurrences. Even in situations where the treatment effects are inconsistent over events, it may still be meaningful to provide estimates of an average effect. In survival analysis situations with multiple endpoints, as with any longitudinal data, special analytical methods are needed to address the dependency of events occurring within subjects. The possibility of censored observations distinguishes multivariate survival analysis from other longitudinal data methods.

In the survival analysis setting, multiple events may represent: 1) occurrences of different types of failure, 2) distinct recurrences of the same type of failure or event, or 3) a combination of 1 and 2. Although our primary interest is in recurrent events of the same type, several important and related techniques have been developed for multiple failures of different types. These are discussed below.

i. Distinct Failures of Different Types

In some situations the occurrence of multiple events which are of different types may be of interest. One example is a placebo controlled

skin cancer trial where lesions of two types could occur (Abu-Libdeh et. al. 1990). In this trial occurrence of one type of lesion did not preclude the occurrence of the other type and each type could occur more than once. To analyze multiple recurrent events of different types, Abu-Libdeh et. al. introduced a mixed poisson regression model with fixed covariate effects and random subject effects which was an extension of a model first introduced by Lawless (1987). This method assumes a parametric form for the baseline hazard as well as for the random effects component. It is this assumed distribution of the random effects component that defines the structure of the dependency among events.

Using a counting process approach and Martingale theory, Pepe (1991) developed an extension of the nonparametric weighted Kaplan-Meier statistic (Pepe and Fleming 1989) to a multivariate setting for distinct event types. To illustrate the method, one example cited in this paper was a bone marrow transplant study in leukemia patients where outcomes included chronic graft versus host disease (GVHD), recurrence of malignancy or death. Each event could occur only once but each patient could experience more than one event. For event k, other events were defined as a) competing risks (occurrence precludes subsequent development of event k, and can be dependent on k) b) censoring events (occurrence precludes observation but not occurrence of k) or c) incidental (all other events). Because of the possibility of dependency among competing risk events, Kaplan-Meier estimates for each endpoint were not appropriate. Instead, Pepe constructed conditional probabilities of events occurring that were functions of several Kaplan-Meier and cumulative incidence estimates which acknowledge the competing risks.

Pepe developed asymptotic distribution theory for these estimators that were made up of several simple functions: survival, cumulative incidence

or cumulative hazard. A moment estimator was used for the variance-covariance matrix and no assumptions were made regarding the dependency among multiple times to event. The resulting estimate of the marginal probability of the event having occurred by time t was a conditional probability that acknowledges the competing risks. A two-sample test statistic was proposed, but there was no direct way to include additional covariates in the method.

ii. Recurrent Events of the Same Type

As noted above the primary interest of this research involves recurrent events of a single type. Although the focus will be on events where the exact time of occurrence is known, there are many situations where the only information available is the total number of events corresponding to intervals between successive observation times. When the time intervals are the same for all patients, this situation is analogous to grouped survival methods such as life table methods in the univariate case. Methods addressing this type of data when the responses are grouped are summarized below.

a. Interval Count Data

In some applications where the event of interest is grouped within intervals, patients may be censored and interval lengths, although usually made up of scheduled visits, may not be uniform for all subjects. Thus, counts between patients are not directly comparable because intervals over which the counts are obtained may differ. In these situations a recurrence rate is often the summary measure of interest, but a simple poisson process model is often inadequate since it assumes events occur independently and at the same rate.

Intersubject variability is likely to cause over-dispersion and it is unlikely that successive events within an individual are independent.

Thall (1988) presented a parametric poisson regression approach for interval count data to address whether time-dependent event rates differ between treatments. A placebo-controlled study of the reported incidence of dyspepsia at increasing intervals [1, 2, 3, 6, 9 months etc...) over a 2 year period was used to illustrate the method. Following a general linear models framework with multivariate response, mixed poisson regression was used that incorporated a multiplicative variance component in the form of random subject effects. Mixing the random effects accounts for within subject dependency, unequal interval lengths and extra-poisson variation. As with any parametric regression, the method is able to incorporate covariate effects. A functional form for the random effects or mixing component must be assumed and this distribution determines the structure of the within subject dependence.

Thall and Lachin (1988) considered a nonparametric method for analyzing random interval count data. The motivating example was a study in which gallstone patients reported the number of episodes of nausea over a series of visits. Intervals on which the corresponding counts were based were random and patients could be censored. The method consists of a multivariate rank test of treatment differences in recurrence rates and was derived from a vector of Wilcoxon-like rank statistics (Wei and Lachin 1984). This approach accounts for censoring and random intervals as well as for dependency resulting from repeated measurements. Although no structure is imposed on the within subject dependency, there is no easy way to adjust for additional covariates.

Freedman et. al. (1989) developed a nonparametric approach to test for the difference between two treatments in bladder tumor recurrence rates. Since tumors were detected at scheduled cystoscopy exams, the exact time of tumor was not known. In addition, more than one tumor could be detected at a single exam. As an overall summary statistic the recurrence rate was defined as the total number of recurrences in the kth group, divided by the total length of observation in the kth group. The authors argued that even though the rate is not constant over time, this estimator provides a summary of the average rate. They claimed that a simple poisson process was not appropriate for this type of data and developed a randomization test for the differences in average recurrence rate. They also provided a bootstrap-based confidence interval for the ratio of the two recurrence rates.

Although no assumptions must be made with respect to within subject dependence of successive event times, this method does require scheduled visits. Another drawback is that adjustment for additional covariates is not easily implemented.

Using the same bladder tumor trial, Cheuvart (1988) proposed a generalized linear model with a logit linear link between recurrence probabilities and covariates. The dependence among recurrences within subjects was assumed to follow a Markov process and transition probabilities were the actual parameters. The Markov states were the number of occurrences during an interval.

Cheuvart used GLM theory to obtain MLE of parameters and the asymptotic variance covariance matrix. Although dependency within subjects was assumed to be Markov, the Markov process was

nonhomogeneous in that transition probabilities could differ for different covariates. The Markov process is a one stage dependence process where, given the first through i-1 events, the i^{th} event depends on the i-1 event only; *i.e.*, past and future are independent given the present (McCullagh and Nelder 1989). The method requires regularly scheduled visits; however, the likelihood is adapted to account for missed visits and right censoring.

b. Exact Time to Event

The remainder of this report will concentrate on the multivariate survival situation when the exact time to an event is known. For this type of data two broad categories of models exist: 1) models that specify the structure of the dependence between events by conditioning on past events or by including random effects in the model, and 2) models that treat the dependence within subjects as a nuisance factor and develop robust covariance estimators that take into account the within subject dependence without specifying its structure. A subset of the models discussed here was evaluated through simulations and examples; the results will be presented in Chapters III and IV.

1. Methods Specifying Structure of Dependency

As mentioned above, there are two general classes of models which require specifying the structure or form of the dependency among events within subjects. These include 1) models which condition on past history and 2) models which include a variance component or random effect for the subjects. In both cases event times within subjects are assumed to be independent conditional

on the additional information or distribution.

Models that condition on past events or past history are often formulated in terms of the multiplicative intensity model for counting processes. A stochastic process $\mathbf{N} = (N_1(t), \ldots, N_k(t))$, $t \in [0,1]$ (can be extended to $[0, \infty]$) is a k-dimensional counting process if each of the k component processes has a sample function which is a right continuous step function with a finite number of jumps at the same time (Andersen, Borgan, Gill and Keiding 1982). In other words, each process N_i counts the events of a *point process*: a random countable collection of points on the real line (Aalen 1978). In our applications, i indexes a subject, and N_i counts the recurrent events for subject i.

The filtration, F_t, is important to the multiplicative intensity model, representing the statistical history or the collection of all events observed on the interval [0, t]. The intensity process of **N** is defined by

$$I_i(t) = \lim_{h \to 0} \frac{1}{h} E(N_i(t+h) - N_i(t) \mid F_t) \quad i = 1, \ldots, k$$

and the multiplicative intensity model is represented by

$$I_i(t) = a_i(t)\, Y_i(t), i = 1, \ldots, k$$

where $\mathbf{a} = (a_1, \ldots, a_k)$ is an unknown function and $\mathbf{Y} = (Y_1, \ldots, Y_k)$ is a stochastic process adapted to F_t. If $I(t) = \delta A(t)/\delta t$ then

$$A_i(t) = \int_0^t a_i(s)\, Y_i(s) ds, \quad i = 1, \ldots, k$$

is called the cumulative intensity process and is also known as the compensator of the process N(t) with respect to the filtration, F_t. The process **Y**, and thus, **A** must be adapted to or predictable with respect to F_t. A process is predictable if its behavior at time t is determined by its behavior on [0,t), for any t (Fleming and Harrington 1992).

The compensator A(t) centers the counting process N(t) around zero, $E(M(t)) = E(N(t)) - E(A(t)) = 0$, thus creating a Martingale with respect to the filtration (Fleming and Harrington 1992). In order for M(t) to be a martingale, N(t) must be an increasing function and A(t) must be a predictable process with respect to the filtration. Martingales are a special type of integral whose properties lend themselves to straightforward proofs of asymptotic results for estimates and inferences based on the multiplicative intensity models in general, and censored survival models in particular (Fleming and Harrington 1992).

Aalen (1978) first developed the multiplicative intensity model approach for right-censored univariate survival data by letting the intensity process $I_i(s) = \lambda_0(s)g(\mathbf{Z})Y_i(s)$. The baseline hazard function λ_0 is arbitrary and g is determined by **Z**, a vector of covariates. The process $N_i(t) = I(X_i \leq t, \delta_i = 1)$ indicates whether subject i had an event at time t and $Y_i(t)$ indicates whether subject i is at risk at time t: $Y_i(t) = I(X_i \geq t)$ with respect to F_t. Since the risk indicator Y is adapted to a filtration it is not determined by the failure process alone, N(t), but can depend arbitrarily on other past events, thus allowing for censoring.

The information in the usual survival data pair (X_i, δ_i) defined in section II. A. i. is the same as the information contained in (N_i, Y_i). The multiplicative intensity model also allows recurrent events by letting $Y_i(t) = 1$ even after an observed failure and $N_i(t)$ be the number of events occurring over time (Fleming and Harrington 1991)

Andersen and Gill (1982) [AG] extended Aalen's general multivariate counting process formulation for survival data to Cox's Proportional hazards survival model by defining the intensity process as:

$$I_i(t) = Y_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}' \mathbf{z}_i(t))$$

AG modeled the intensity rate function conditional on the relevant event, censoring and covariate history up to time t, where the relevant history is included in the vector of model covariates. As with the Cox model, the intensity process I(t) is related to the covariates through the function $\exp(\boldsymbol{\beta}'\mathbf{z}(t))$ and the baseline hazard function is arbitrary. The AG model differs from the Cox model in that it can include recurrent events. Essentially the method sums Cox's likelihood over all events.

AG developed three specific models which they applied to the data from a study of admissions to psychiatric hospitals among women giving birth during a given time period. Let $N_i(t)$ be the number of admissions for woman i in the relevant time interval and $Y_i(t) = 0$ if woman i is a resident of a hospital at time t, $Y_i(t) = 1$ otherwise. $N_i(t)$ is thus a counting process with the

intensity function given above. In their first model AG assumed that a woman's intensity for admission at time t was influenced by the woman's parity and time since date of birth of the child, but not by any pattern of previous admissions. Their second model also included the woman's age in the covariate vector. Finally, a third model added a time dependent indicator variable for whether the woman had been admitted to a psychiatric hospital within the 30 days prior to time t. The third model is thus semi-Markov, since the probability of being admitted depends on the time since last admission. The first two models are Markov: The transition probabilities for going from not admitted to admitted depend only on the state the woman is in at time t and other covariates, but not on prior admissions history.

Estimates of covariate effects and inference are based on partial likelihood information. Estimates of covariates are shown to be asymptotically normal and consistent. The covariance matrix is obtained from the sample information matrix. The Wald and likelihood ratio tests that are used to make inferences about the model parameters are based on the partial likelihood principle and require that the models and correlation structure be correctly specified in order for the tests to be valid.

In the AG model the pattern (number, timing and order) of previous psychiatric admissions is not considered a part of the medical history that is relevant to the intensity for admissions at time t. The exception is model 3 where the way the occurrence of a previous admissions is considered relevant is whether or not a woman had at least one prior admission within the past 30

days. The risk set for the k*th* admission is not restricted to patients who have had k-1 admissions but includes all patients who have either not been censored or are not in hospital at time t. Treatment effect and other covariate effects are estimated over all recurrences and are not allowed to vary among the recurrences.

Another important characteristic of the AG models is that they all assume the baseline intensity function is constant over all events, which is probably an unrealistic assumption. It may be years before patients suffer their first MI, but once they've had an MI, physiological changes may occur that make the occurrence of a second MI more likely. Even using AG's own example of postpartum admissions to psychiatric hospital: It seems more reasonable to assume that once a woman has been admitted to a psychiatric hospital, her risk for subsequent hospital admission increases, instead of staying the same. Unless the assumption of a baseline hazard that is constant over event number is reasonable for a specific example, the use of this method may not be appropriate.

Finally it is important to emphasize that in these models, any relationships of the intensity function to past events must be explicitly included in the model as covariates; i.e., the dependence structure is specified by the model. Therefore, inferences about the model parameters are not valid unless the model has been correctly specified. Even model 3, which attempts to account for prior event time through the use of covariates, may not account properly for dependence since the way the relationship of event time to previous event time is

modeled may not reflect the true relationship.

Another drawback to this model is that the probability of admissions at time t depends on total time since the start of the study, which actually introduces correlations among the admissions times into the model. If the first event time relative to the start of the study is short, the second event time is also likely to be short, but if the first event time is long, then event times for all subsequent events will also be long. Although correlations of this type were introduced into the models they were not accounted for at all in the estimate of the covariance matrix for the first two Markov models. For the third semi-Markov model, it is unlikely that the correlations introduced through modelling total event times will be adequately adjusted for by the inclusion of time-dependent covariates.

Prentice, Williams and Peterson (1981) (PWP) used a slightly different multiplicative intensity model for survival data with multiple recurrences. They demonstrated their methods with data comparing the occurrences of serious infections in post bone marrow transplant patients with and without graft versus host disease. PWP modeled the intensity function for an individual, which at time t is defined as the instantaneous risk of an event conditional on the entire event, censoring and covariate history before t using an extension of the stratified Cox model. Stratum (s) was defined by event number (k), $s = k + 1$ and all patients are in stratum one until they experience their first event, at which time they enter stratum two; etc. Thus the model differs from AG in that subjects are not at risk for the k^{th} event unless they have experienced k-1 events. This implies that the ordering of the

events is considered meaningful and conditioning on past history, or the filtration, includes the number of previous events in addition to the covariate and censoring information.

The stratification structure results in another important difference from AG: The model allows covariates and hazard functions to vary from event to event, and as with any stratified Cox model, the proportional hazard assumption does not apply to the stratification variable. Estimates of covariate effects included in the model, such as treatment, can be obtained either for individual strata or "pooled" across strata. The pooled treatment effect, and any other covariate effects, are adjusted for the stratification variable but the model makes no assumptions about the effect of the stratification variable on event times.

PWP developed two stratified Cox models. In the first model (which they called model 2) the baseline intensity function for infections depends arbitrarily on the total time in the study as well as the number of previous infections (strata). This model is Markovian, since the rate does not depend on the timing of the previous events. Let each of i subjects experience up to k recurrences of infection with strata indexed by s. The intensity function for model 2 (PWP_2) is:

$$I_{is}(t) = Y_{is}(t)\lambda_{0s}(t)\exp(\beta' z_{is}(t)).$$

Their second model was a Semi-Markov model which they called model 3 (PWP_3). For this model, in addition to the number of previous tumors, the intensity function depends arbitrarily on time

since last infection, or gap time between successive tumors, instead of total time in the study. The intensity process is defined as:

$$I_{is}(t) = Y_{is}(t)\lambda_{0s}(t - t_{s-1})\exp(\beta' z_{is}(t)).$$

Both models make assumptions about the within subject dependence of the recurrence times. PWP_2 assumes that the hazard function of the kth event depends on the total time in the study but is independent of the previous k-1 recurrence times. By allowing the hazard function to depend on total time in the study, PWP_2 (like the Anderson-Gill procedure) probably introduces correlations among event times within an individual without accounting for this relationship in the estimation of the covariance matrix. PWP_3 assumes that the hazard function of the k^{th} recurrence depends on the time of the $(k-1)^{th}$ occurrence but on no other. Although this model doesn't introduce correlations between event times, it doesn't adjust for them if they exist. Estimates of covariate effects, both for individual strata and "pooled" over strata, are based on maximum partial likelihood and standard error estimates are from the second derivative of log partial likelihoods (Fisher's information). Generalization of the score statistics provides a generalization of the log- rank statistic.

Thus, as with the AG models, inference about the PWP model parameters is valid only if the dependence structure has been correctly specified.

In simulation studies, Wei Lin and Weissfeld (1989) showed that

the PWP models are sensitive to misspecification of the model. When the dependence structures of events within subjects differ from those specified in the model, the empirical alpha levels for the Wald tests of no difference between two treatments exceeded the nominal levels. In fact, for the Markov model, the empirical alpha exceeded the nominal level in all cases, while the semi-Markov model performed badly when correlation between event times increased.

Pepe and Cai (1993) also showed that the size of the log rank test for treatment effects for recurrent events obtained from the PWP Markov model exceeded nominal levels when models were misspecified. The Markov model performed badly when event times were generated using semi-Markov rather than Markov assumptions.

Another approach to multiple occurrence data that also uses Cox's proportional hazards and partial likelihood principles turns out to be a special 2-sample case of PWP (Wei Lin and Weissfeld 1989). Gail, Santner and Brown (1980) modelled time to mammary tumors in rats using Markov and semi-Markov assumptions. Unlike the other examples using tumor data described in section II.C.ii.a., tumor times were not considered grouped; *i.e.*, occurring at some unknown time within an interval. Rats were palpitated daily and tumors found on a particular day were attributed to that day.

Gail, et. al. developed several different models, including a Cox-like semi-parametric model with unspecified hazard function, and a parametric model assuming a Weibull baseline hazard. Models

were developed for responses defined as both gap times between successive tumor occurrences, similar to PWP_3, and time from study entry to each tumor occurrence, similar to PWP_2 . All hazard models conditioned on either time of last tumor (semi-Markov) or tumor number and time since study entry (Markov). Thus, the dependency structure was specified and, as with the PWP and AG models, correlations are either introduced and not accounted for or not accounted for if they exist.

Using the same rat tumor data as Gail, Santner and Brown, Lawless (1987) developed several poisson models with random subject effects to account for extra poisson variation. Lawless assumed proportional intensity and a poisson process; however, unlike the following frailty models, the random subject effects do not account for the serial dependence of event times within an individual. His models are somewhat similar to the AG approach in that the hazard for the k*th* recurrence does not depend on the previous k-1 occurrence times.

Another area of research which is applicable to the multivariate survival data is the frailty model. These models use a random effect or variance component to account for the dependence within subjects. Observed survival times within an individual depend on the same random unobservable variable, the frailty, via the proportional hazards model. Thus for subject i:

$$\lambda_{ik}(t;Z_{ik},w_i)=w_i\lambda_0(t)e^{\beta' z_{ik}(t)}.$$

The frailty is the same for hazard functions from the same subject and the survival times are assumed independent conditional on

the frailty variables. Oakes (1989) discussed a bivariate frailty model while Clayton and Cuzick (1985) examined more general multivariate frailty models based on the gamma distribution. The statistical analyses for these models are complicated and asymptotic results are difficult to obtain. Hougaard (1986) developed a method for frailty models where the frailty has a positive stable distribution. Although these models are computationally simpler, they are unrealistic since the expected value of the distribution is infinity. Furthermore, the statistical properties of the inference based on these procedures have not been developed (Lee, Wei and Amato 1993).

2. Minimal Assumptions with Respect to Dependency Structure

Instead of trying to specify the structure of the dependency of failure times within a subject by conditioning on past events or using random effects, the models in this section treat the dependence of recurrent survival times within patients as a nuisance factor.

Wei, Lin and Weissfeld (1989) (WLW) proposed a model that has the important advantage of making no assumptions with respect to the dependency structure of recurrent events. They formulate a Cox-type proportional hazards model for the marginal failure time distributions of each type of failure. The following hazard is defined for the k^{th} event of the i^{th} subject:

$$\lambda_{ki}(t) = \lambda_{k0}(t)\exp\{\beta_k' Z_{ki}(t)\}.$$

As with other models based on the proportional hazard approach, the baseline hazard is arbitrary; and, as with the PWP approaches, the baseline hazard is permitted to vary among failure events. The vector of regression parameter estimates must also vary from event to event. Failure specific regression parameters are estimated for the k events using the ordinary maximum partial likelihood methods for each marginal model. This approach to estimating the covariate effects assumes that the k event times are independent; however, the estimates are unaffected by this assumption and the $\hat{\beta}_k$ are shown to be consistent and asymptotically normal. The covariance matrix, **Q**, is not based on the Fisher information matrix which requires the independence assumption, but on a robust variance estimator that accounts for the correlation (Lin and Wei 1989). The joint distribution of the estimates of K covariate effects is shown to be asymptotically normal. WLW also provide a linear combination of covariate effects over the k failure events to estimate an "average" effect over the k events. This estimator has been shown to have the smallest asymptotic variance among all linear estimators (Wei and Johnson, 1985). Let $\eta = \beta_1$ be the vector of covariates for a particular effect over the k events. For example, η_k might be represent the effect of treatment for the k^{th} event. Suppose that $\eta_1 = \ldots = \eta_k = \eta$ and $\hat{\Psi}$ is the estimator of the covariance matrix for $(\eta_1 = \ldots = \eta_k)'$, which is obtained by selecting appropriate rows and columns from $\hat{Q}$. Then WLW estimate η by a linear combination of the $\hat{\eta}_k$:

$$\sum_{k=1}^{K} c_k \eta_k, \quad where \sum_{k=1}^{K} c_k = 1. \tag{2}$$

and

$$c=(c_1,\ldots,c_k)'=(e'\hat{\Psi}^{-1}e)^{-1}\hat{\Psi}^{-1}e, \quad where \quad e=(1,..,1)' \tag{3}$$

The standard error of $\hat{\eta}$ is estimated by:

$$se(\hat{\eta})=\sqrt{c'\hat{\Psi}c} \tag{4}$$

Thus, with the Wei, Lin and Weissfeld procedure inference is possible for individual events as well as globally over all k events.

So WLW provide the estimator for $\eta_1 = \ldots = \eta_k = \eta$; however they claim that even if the individual estimates are not equal they can still be combined to draw conclusions about the "average effect" as long as there are no qualitative differences among the individual estimates. This is analogous to the situation of investigator by treatment interaction in a multicenter clinical trial. If quantitative differences exist among investigators, *i.e.* treatment effects differ but are in the same direction, the treatment effects are still combined and an overall estimate is obtained. If treatment effect differences are in different directions, overall event effects may not make sense. This emphasizes the need to be able to investigate the individual event estimates to see that they are homogenous or at least differ only qualitatively before they are combined into an overall estimate.

The major drawback of the Wei, Lin and Weissfeld approach is that since it is a marginal approach, the hazard function for each of k events at time t is estimated using a risk set that includes all patients who have not been censored or experienced the k^{th} event prior to time t. That is, the risk set for the kth event includes subjects who have not yet experienced event k-1. If the ordering of the outcomes is meaningful then the WLW risk set makes the interpretation of the relative hazard ratio difficult for the individual events: multiplicative effect of treatment relative to placebo on the hazard rate for the kth event, where not everyone has experienced k-1 events? Besides making the interpretation of the covariate parameters difficult, Pepe and Cai (1993) point out that the survival functions are likely to be correlated since a low cumulative marginal hazard for the k^{th} event time implies a low cumulative event time for the $(k+1)^{st}$ event time. In other words, for this particular risk set, information from the first event is included in estimating parameters for subsequent events.

Lee, Wei and Amato (1992) proposed a marginal Cox proportional hazards model similar to WLW's. They estimate parameters based on assumptions of independence and then use the robust variance estimate due to Lin and Wei (1989). The model is considered marginal in that all subjects who have not been censored prior to time t are included in the risk set for each failure, regardless of the number of previous failures. Their approach differs from WLW in that their model uses the same nuisance baseline hazard for all failures, whereas in the WLW model the nuisance baseline hazard is allowed to vary with each failure (Lin 1993). The model differs from AG in that information from each event is included in the risk set.

Another approach that can be based on the marginal model is the accelerated failure time model. In these models, if $\beta' z < 0$ then the covariate is accelerating the time to failure. As an example of this type of model, Lin and Wei (1992) applied the WLW approach using linear regression to relate the logarithm of marginal failure times to covariates (Lin 1993).

Pepe and Cai (PC) developed a model for recurrent events that is based on rate functions. Their method differs from the conditional multiplicative intensity process in that instead of modelling the instantaneous intensity at time t conditional on the entire event, censoring and covariate history up to time t, they model the intensity for an event k that is conditional only on being at risk at time t and having had k-1 previous events. Their model is similar to the PWP_2 model except that it uses a robust variance estimator that does not depend on specifying the within patient dependence structure of events but accounts for correlation between event times. In other words they developed a variance estimate that is not dependent on the assumptions of the Markov model, as it is with PWP_2. The model differs from the WLW marginal model in that patients are considered at risk for the k^{th} event only after they have experienced k-1 events; however, their risk set also differs from the PWP restricted risk set in that to be in the risk set patients must have had k-1 events before time t, the event time of the kth event. Thus subject i is not considered at risk for a third event if his second event has not occurred prior to the time of the third event in question. Since PC model total time since study start, this may be the most appropriate risk set;

however, it will tend to greatly reduce risk set sizes at later events.

Pepe and Cai proposed the following Cox-type model for the recurrence rate of the kth event:

$$r_{k|(k-1)}(t) = r_{k0}(t)\exp\{\beta_k' z\},$$

where r_{k0} is an arbitrary baseline rate function for the kth event. The authors' notation for the rate function helps to differentiate it from the marginal hazard function of WLW, $\lambda(t)$, and intensity functions of PWP, i(t). In all three methods the functions are the same for the first occurrence of the event but differ for recurrent events depending on risk set and model assumptions.

The authors proposed estimating the regression parameters for the kth occurrence using the following estimating equation which was motivated by the general estimating equations idea of Liang and Zeger (1986). The equation uses counting process notation:

$$S(\beta)=\sum_{i=1}^{n}\int_0^t z_i(t)\{dN_i(t)-Y_i(t)\exp(\beta' z(t))\,\hat{r}_0(t)dt\}.$$

Since r(t) is unknown it is replaced with the Nelson estimator (Fleming and Harrington 1992). For the kth event the definition of r(t) is:

$$\hat{r}_{0k}(t)dt=\sum dN_{ik}(t)/\sum Y_{ik}(t)\exp\{\beta_k' z_i(t)\}.$$

The above estimating equation for recurrent events is motivated by the partial likelihood score equation for the first event. Furthermore, with a single dichotomous covariate, the estimating function for the kth event is shown to be similar to the log rank statistic in that it is of the form $\sum (O-E)$. After developing the score-like equation and finding an estimator that is a solution to $S(\beta) = 0$, Pepe and Cai developed asymptotic distribution results for the estimator, providing a robust variance estimator for $\hat{\beta}$. Inference based on this variance estimator takes into account correlations of event times within subjects without assuming any structure.

The authors illustrated their methods for recurrent events with a study comparing the effects of two treatments on the rate of infection in bone marrow transplant patients. Although they provide proofs that their methods are applicable to examining and combining treatment effects for each of k recurrent events of the same type they demonstrate the method only by distinguishing between first infection rate, $r_F(t)$ and recurrent infection rate, $r_R(t)$. A log rank type statistic is calculated for treatment comparisons in each type of infection rate, and then the two are combined for an estimate of the overall treatment effect.

For the purposes of this research, the method was extended to situations where the model includes covariates in addition to a dichotomous treatment variable. In this situation inference would be based on score or Wald type tests using the estimated βs and the corresponding rows and columns from the covariance matrix. For the purposes of this research, the PC method was also

extended to modelling the rate function for each of the k events in order to compare their results with other methods. For inference over all k events the linear combination of parameter effects proposed by WLW was used. This approach is shown in equations (2), (3) and (4).

In addition to deriving regression parameter estimates and variance estimates for the recurrent data, Pepe and Cai also provided point estimates and confidence bands for the rate functions based on general estimating equations. They also developed methods to analyze failure data with a single endpoint, but with time varying covariates. Since the emphasis of our research is on estimates of differences in treatment effects for recurrent events and related inference, the additional estimation and single endpoint topics are not reviewed here.

The PC approach has the advantage of using an appropriate risk set for modelling total times (unlike AG, WLW and PWP model 2) without the need to specify the dependence structure. Although the PC method has several attractive features, a possible drawback is that they do model total time rather than gap times. The implications of how event time is defined is discussed in the following section. Because of concern with this aspect of their model, a new approach was investigated that models gap times rather than total times and as a result uses a less restrictive risk set. Patients are at risk for event k if they have had k-1 previous events, without placing a restriction on the time the event occurs. This method will be referred to as the modified Pepe/Cai method, PC_m.

D. Features of Existing Methods Chosen for Evaluation

Based on the literature review presented above, AG, PWP_2, PWP_3, WLW, PC and a modified PC (PC_m) were the six types of models felt to be most appropriate for analyzing the type of recurrent failure data of interest here. These models can be categorized with respect to whether there is a need to specify the dependence structure, whether they model total event time or gap time, whether or not the risk set is restricted to patients having k-1 previous events, whether covariates are constrained or allowed to vary among events and finally, whether the hazard function can vary among events. Table 1 summarizes some of the important features that distinguish the six models. Specific features that may have advantages for analyzing the type of data we are interested in are shown in bold.

Table 1
Model Assumptions and Features

Model	Need to Specify Dependence	Event Time	Risk Set	Constrain Covariates	Constant Hazard
AG	Yes	Total	Unrestricted	All	Yes
PWP_2	Yes	Total	**Restricted**	**Some**	**No**
PWP_3	Yes	**Gap**	**Restricted**	**Some**	**No**
WLW	**No**	Total	Unrestricted	None	**No**
PC	**No**	Total	**Restricted***	None	**No**
PC_m	**No**	**Gap**	**Restricted**	None	**No**

*Risk set restriction for the Pepe-Cai approach is based on Time of event as well as number of previous events.

The following section reviews and evaluates the six approaches with respect to these important model assumptions and features.

i. Model Assumptions and Definitions

a. Dependence

One of the most restrictive aspects of recurrent event failure data is that failure times within an individual may or may not be independent of previous failure times. For example, in the SOLVD treatment trial subjects who were admitted to the hospital for congestive heart failure early in the study may be more likely to be hospitalized a second time sooner relative to their first admission, than patients who were first admitted to the hospital later in the study. As illustrated below in the section on event times, when total time since study start is modelled rather than time between events, correlation between event times is actually introduced into the model.

Furthermore, whether the correlations among event times are real or introduced, we may or may not know the structure of the dependency. In most practical examples we will not know the dependency structure of the event time data in advance. The Andersen and Gill proportional hazard intensity model was one of the first approaches to modelling multiple event failure data using the multiplicative intensity model. With these models it is necessary to specify how past events effect the successive event times within an individual. AG developed a Markov model where time to an event depended on the subject's current state but not past event times, and a semi-Markov model where time to event was assumed to depend on time of last event, in addition to the state the patient is in at the time. These models make explicit assumptions regarding the structure of dependencies within the individual and can actually create correlations between event times. If the assumed structure

is incorrect and correlations exist but are not accounted for in the covariance structure, then inference based on the model is not valid.

PWP introduced several stratified models that were more flexible than the AG model in some respects (discussed below) but still required that dependencies of event times within subjects be modelled. As with AG, their first model allows the probability of an event to depend on time since study start. This introduces correlation between successive event times since a long time to first event implies a long time to subsequent events. Their second model specifies that event times depend on the time since the last event, which avoids the problem of creating correlation among the event times but still does not account for correlations if they exist. The drawbacks of specifying the dependency structure discussed for the AG models also hold for the PWP models.

On the other hand, the methods proposed by WLW, PC and the modified PC do not require specifying the dependency structure of the hazard function on past events. Both methods treat this dependence as a nuisance and develop covariance estimates that account for the correlation between event times without making assumptions about the structure.

It seems appropriate to investigate whether inference based on models that either introduce or don't adjust for existing correlations have appropriate alpha levels under the null hypothesis.

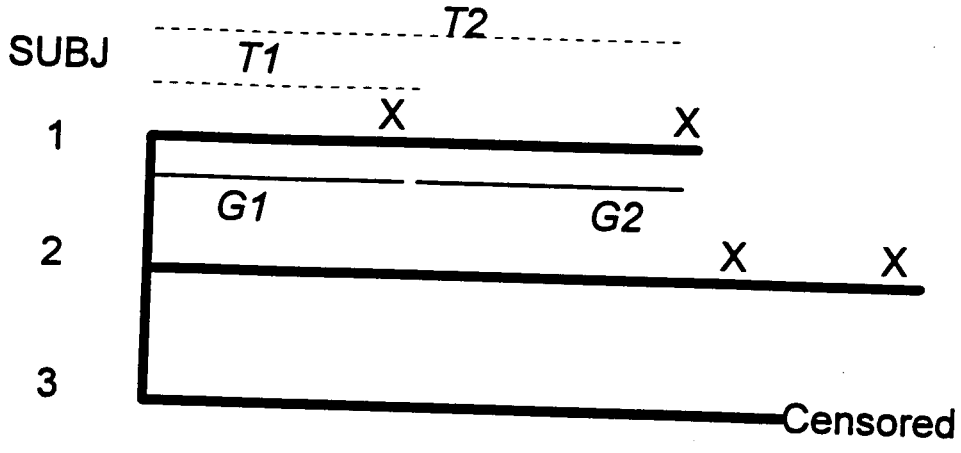
b. Event Times

Another important difference among the models is whether event

time is modelled as time since study start (total time) or as time since last event (gap time). AG, PWP_2, WLW and PC all model total time, while only PWP_3 and the modified PC model gap times. The definition of event time affects the interpretation of the model parameters. For example, for methods that model total time, the parameter estimate associated with a second event represents the relative risk for the second event, calculated from the start of the study. This interpretation may make sense in studies of somewhat independent events such as tumors or infections, where it may be reasonable to assume events are developing simultaneously and that the risk for each event begins at the same time. For example, patients might be at risk of infection by a variety of viruses from the start of the trial; first and second events simply reflect the order of their occurrence. Parameters from methods that model gap times, on the other hand, represent the relative risk of a second event from the time of the first event. This interpretation is more meaningful in examples such as hospital admissions or myocardial infarctions, where the risk for a second event does not begin until after the first event has occurred. For example, a first myocardial infarction alters the cardiac anatomy and a second MI can only be defined in the context of these anatomical changes.

Besides the interpretation of model parameters, the definition of event time can affect the statistical features of the models. The diagram in Figure 4 can be used to illustrate the difference between the two types of event times used by the methods and to point out the drawback of modelling total times as opposed to gap times.

Figure 4
Illustration of Event Times and Risk Sets



RisK Set For Subj 1, Event 2:

AG, WLW: 1, 2, 3
PWP, PC-M: 1,2
PC: 1

The figure summarizes the failure experience of three patients. X's represent events and the bold lines represent time. The first two subjects each have two events and the third is censored before an event occurs. The top two arrows show total times, or time since the start of the study for the first subject: T_1 is time for event 1 and T_2 is time since study start for event 2. G_1 and G_2 below the time line represent the corresponding gap times, or time since last event for the same subject. For the first event, gap time and total time will be equal. From this illustration you can clearly see how total times are introducing correlation, even when the two gap times might be independent. If T_1 is long then T_2 must be long. Another way to describe this problem is to remember that T_2, total time for the second event, is the sum of the 2 gap times, G_1 and G_2, and therefore is the sum of two random variables. Another interpretation that is more appropriate for "total time" type events, such as the occurrence of tumors, is that the total times are order statistics, which of course are correlated.

Thus methods that model time from study start will introduce correlation into analysis; however, some of the methods will account for this by adjusting for the correlation. It will be interesting to investigate the actual significance levels of models with different definitions of event time.

c. Risk Set

Another distinguishing feature among the six models is the risk set that is used in calculating the effects of covariates on hazard rate. Both AG and WLW assume that all subjects are at risk for any failure at time t as long as they are still in the study at time t, regardless of

the number of previous failures. PWP and modified PC both include in the risk set for event k, only those subjects who have experienced k-1 events. PC adds the further restriction that subjects must experience k-1 events before time t, the time of the kth event. The diagram in figure 4 can also be used to illustrate the differences among the three types of risk sets used by the methods. Consider the second event of the first subject. For the unrestricted risk sets of AG and WLW, all 3 subjects are included in the risk set for the first subject's second event. Therefore information from subject 3 will be included in estimating the log hazard ratio for all events, not just the first event. Thus this unrestricted risk set results in first event information being included in subsequent events and can create misleading parameter estimates. Furthermore, as pointed out by Pepe and Cai (1993), if the risk set includes all patients, then the cumulative hazard for the kth event will always be smaller than for the k-1 event and this forced ordering implies correlation among the parameter estimates.

Like the choice of the event time definition, the choice of the appropriate risk set depends partly on the interpretation of the relationship of the recurrent events to one another. In the case of distinct events of different types (*e.g.*, the occurrence of post-transplant infection, organ rejection or death; or the occurrence of tumors in different parts of the body), it may be reasonable to assume that all patients in the study are at risk for all events from the beginning of the study. However, if a single event type is considered recurrent and the ordering of events is considered meaningful (e.g., the occurrence of first, second and third myocardial infarctions) then the risk set for event k should contain only those patients who have had k-1 events, otherwise it is difficult to interpret

the event specific regression parameters.

In our application, events are considered ordered and patients should not be at risk for an event k, unless they have experienced k-1 events.

For the restricted risk set defined by both PWP models and the PC_m method, subjects 1 and 2 are at risk for subject 1's second event, because each has had a prior event, but subject 3 is not. Since Pepe and Cai make the further restriction that to be at risk for a second event, a subject's first event must have occurred at a time prior to subject 1's second event, subject 2 would no longer be considered at risk because his first event does not occur until after subject 1's second event. According to the PC model, the only person at risk for subject 1's second event is subject 1. This definition of the risk set makes the most sense when modelling total time, but is not necessary if modelling gap times. An important drawback of PCs restricted risk set is that it will tend to greatly reduce risk sets at later events, and could therefore reduce the power to detect important covariate effects.

Regardless of the impact of the choice of the risk set on the interpretation of the hazard function or parameter estimates, it would be interesting to compare the performance of models that use different risk sets and determine whether the numerical results are meaningful. The significance levels and power of the WLW method in different situations can be evaluated, even if the parameter estimates are questionable.

d. Interpretation and Estimation of Covariate Effects

Each of the six methods considered can be thought of as an extension of the proportional hazards regression model, in which case the effects of covariates on the hazard rate can be represented by a vector (single event/average over multiple events) or matrix (multiple events) of betas where each parameter represents the difference in log hazard per unit change in the covariate. Thus for a single dichotomous covariate, such as treatment, beta (β) represents the change in the log hazard rate relative to the reference or baseline hazard due to one treatment and exp(β) represents the multiplicative change in the hazard rate, or the relative risk. For continuous covariates such as age or blood pressure, β represents the increase or decrease in log hazard relative to the per unit change in the covariate.

Rather than merely providing a p-value as evidence of a meaningful covariate effect, presentation of the parameter estimates along with standard errors or confidence intervals gives the researcher an idea of the extent of the covariates effect on the baseline or reference hazard. This provides an estimate of how large a decrease or increase in hazard rate we might expect and whether it is meaningful, regardless of whether it is significant.

The definition of the risk set and whether event time is calculated as total time or gap time both influence the interpretation of the survival functions, covariate effects and the parameter estimates themselves. This issue is illustrated by the Kaplan-Meir estimates of patients hospitalized for any cause during the SOLVD treatment trial (SOLVD 1991). The details of this study are discussed in Chapter 3.

Figure 5 provides the estimates for proportions of patients hospitalized for the first, second, third and fourth events for a group of patients receiving enalapril and for a group receiving placebo. For each event number, the difference between the curves of the two treatment groups is a manifestation of the log hazard ratio. In this illustration, the time to an event is calculated relative to study start and the risk set for each event includes all subjects who remained in the study at the time of the event. Thus, for the second event, the Kaplan-Meir curve for each treatment group represents the proportion of all patients randomized who experienced a second hospitalization. For example, 31.9% of the enalapril group and 39.7% of the placebo group were hospitalized twice within the first 24 months. The interpretation for the third and forth events is similar. Based on this graph, the difference between the placebo and enalapril treatment groups that is apparent at the first event appears to be sustained over all events. This illustration represents the WLW approach and demonstrates how this method would result in estimates of treatment effect that are of a similar size for each of the four events, even when the treatment effect actually differs among the events.

If gap times are used instead of total times (*i.e.* time from first hospitalization to the second, time from second hospitalization to third and so forth) and the risk set is restricted to include only those patients who have experienced the previous event, the results are very different. Figure 6 gives the curves for the time to first event, and of course these are the same for both definitions. But as seen in Figure 7, the treatment effect is diminished by the second event and virtually nonexistent for the third and fourth events (Figures 8 and 9). Thus, although enalapril may be protective against the first event, once the first event has occurred there appears to be little additional

benefit. This last set of illustrations represent the PWP_3 and PC_m approach to modeling multiple event survival data.

Figure 5

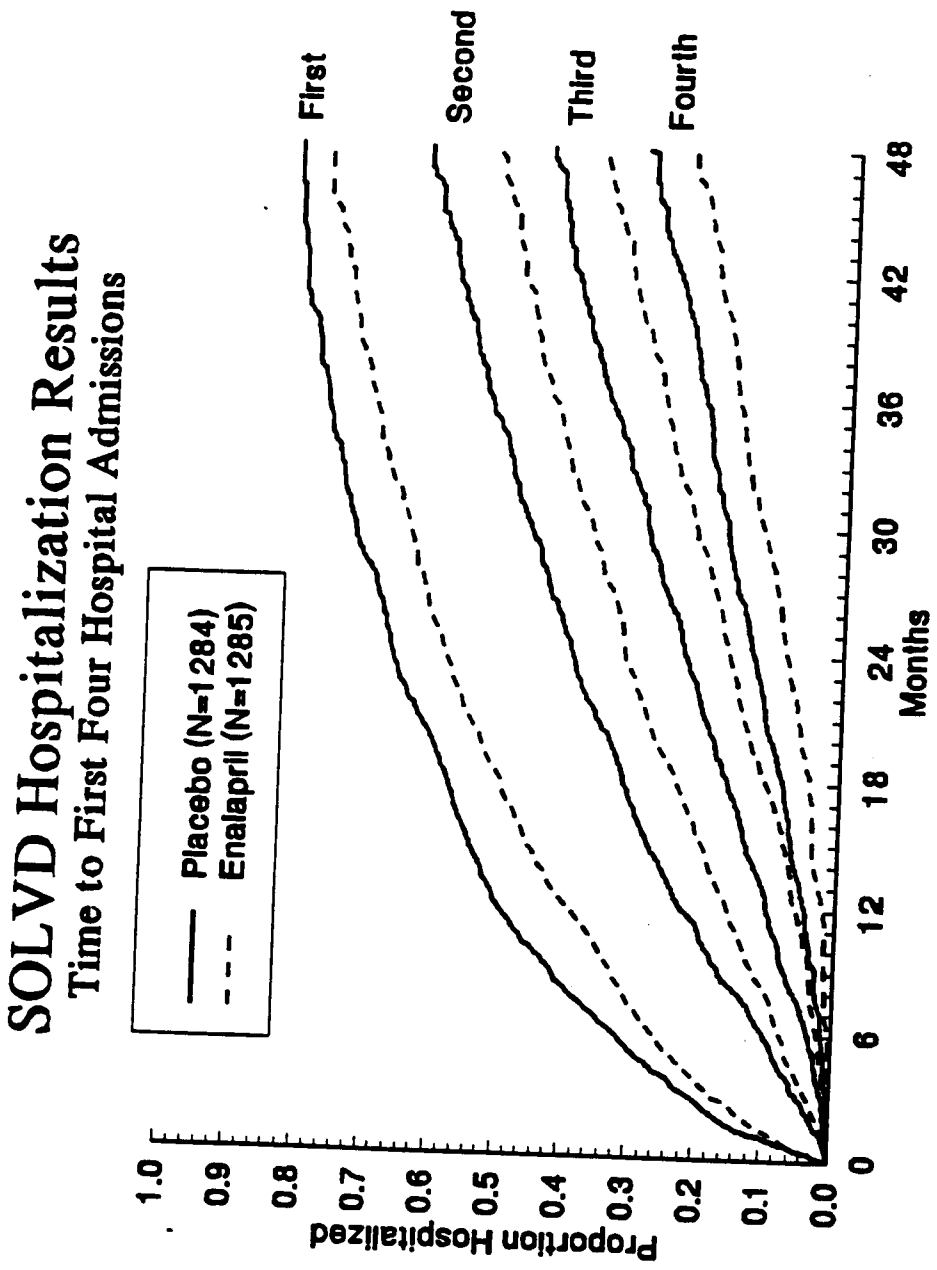

SOLVD Hospitalization Results
Time to First Four Hospital Admissions
Placebo (N=1284)
Enalapril (N=1285)
First
Second
Third
Fourth
Proportion Hospitalized
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
6
12
18
24
30
36
42
48
Months

Figure 6

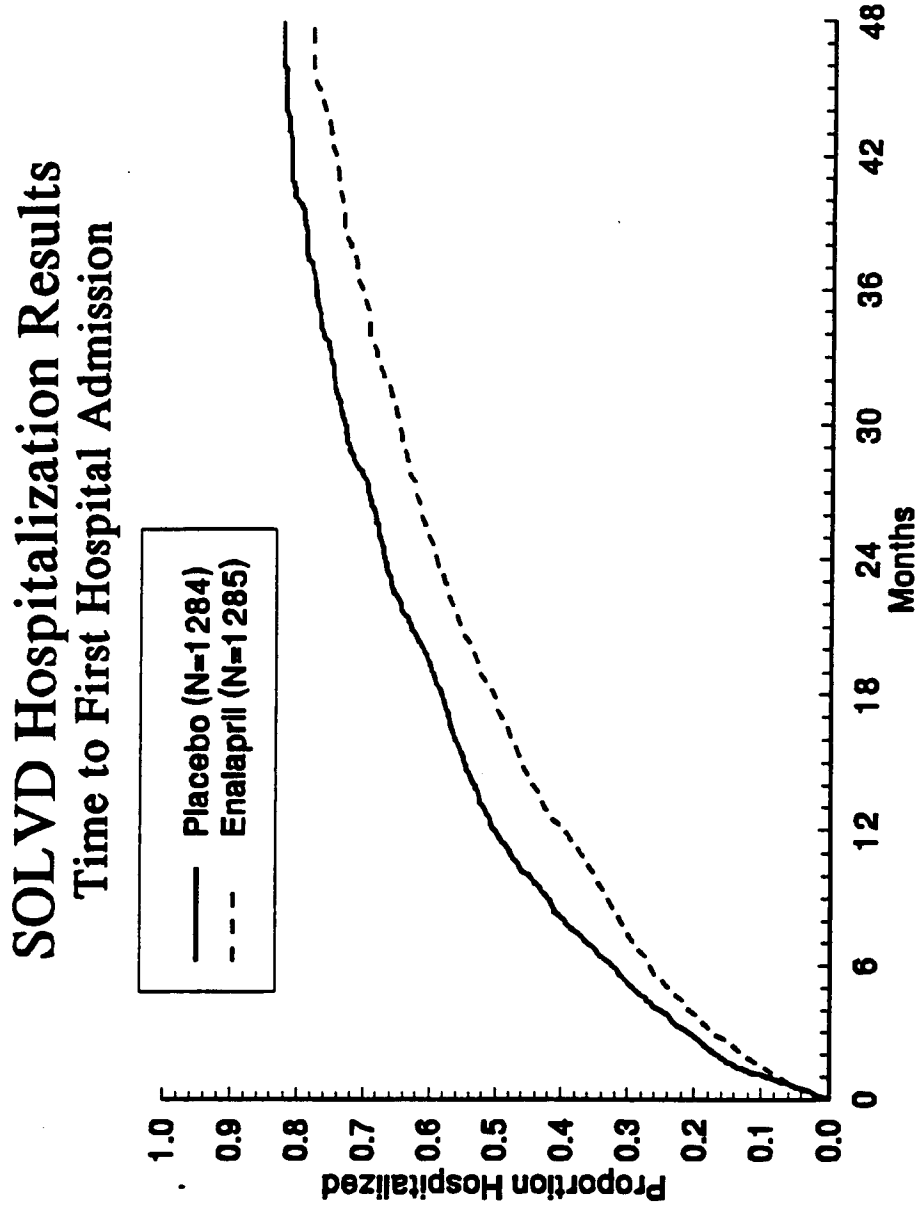

SOLVD Hospitalization Results
Time to First Hospital Admission
Placebo (N=1284)
Enalapril (N=1285)
Proportion Hospitalized
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
6
12
18
24
30
36
42
48
Months

Figure 7

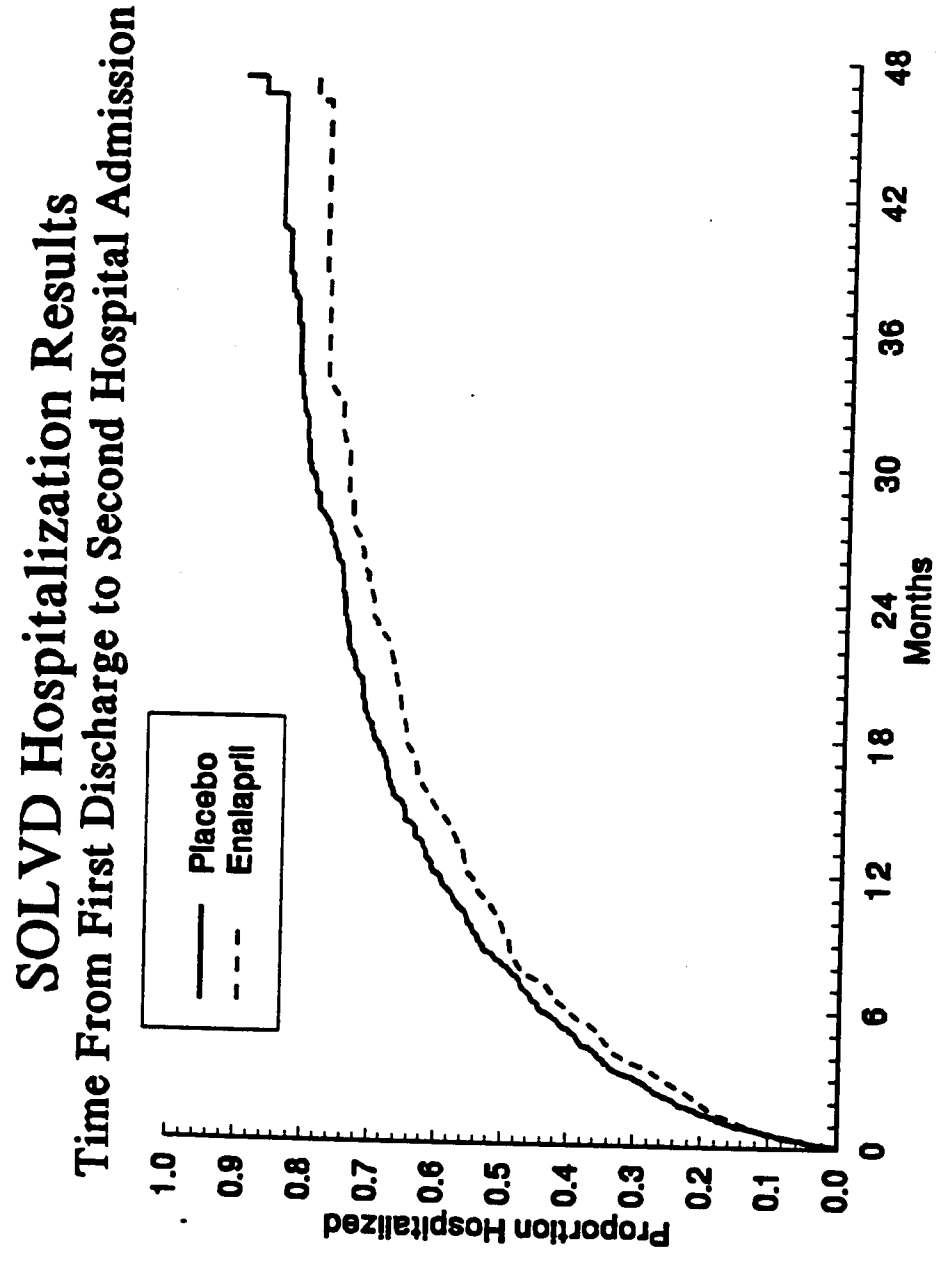

SOLVD Hospitalization Results
Time From First Discharge to Second Hospital Admission
Placebo
Enalapril
Proportion Hospitalized
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
6
12
18
24
30
36
42
48
Months

Figure 8

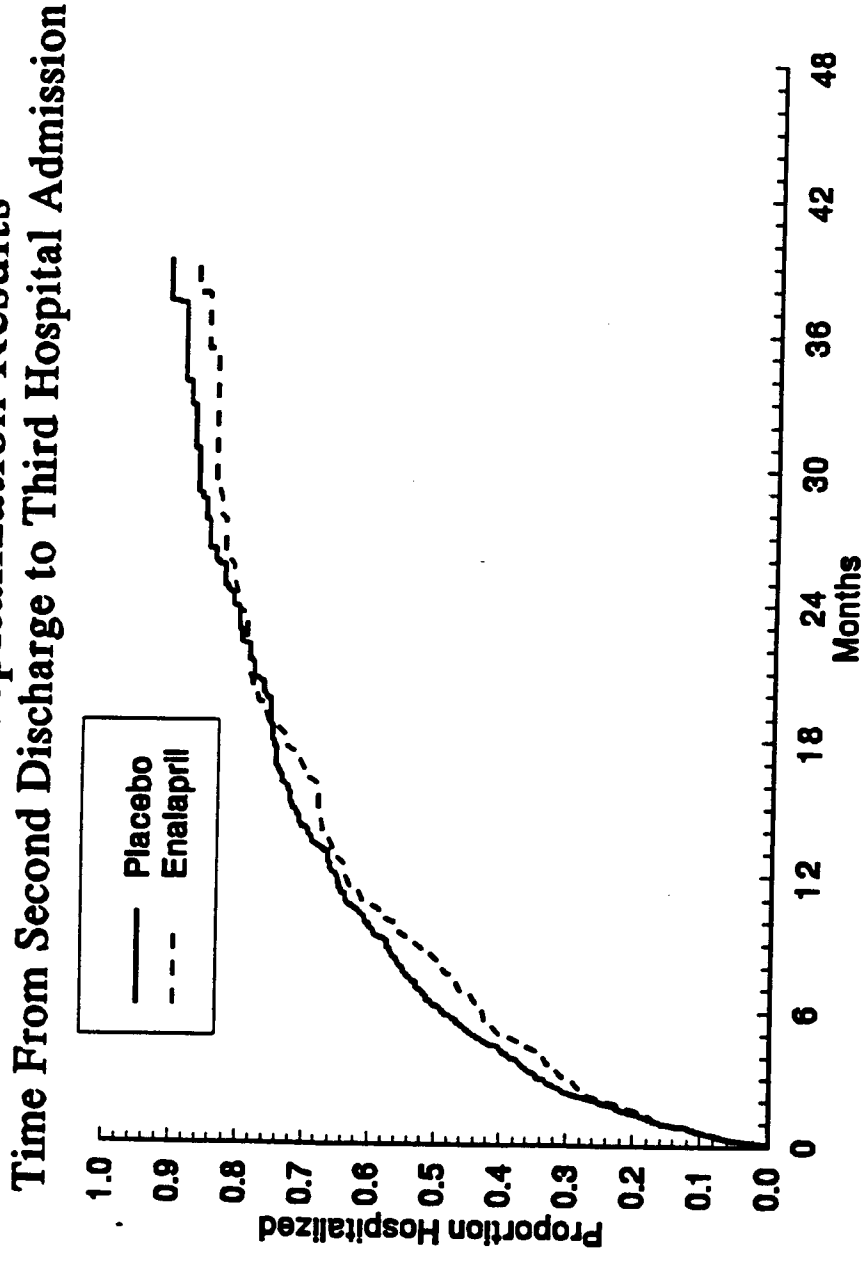
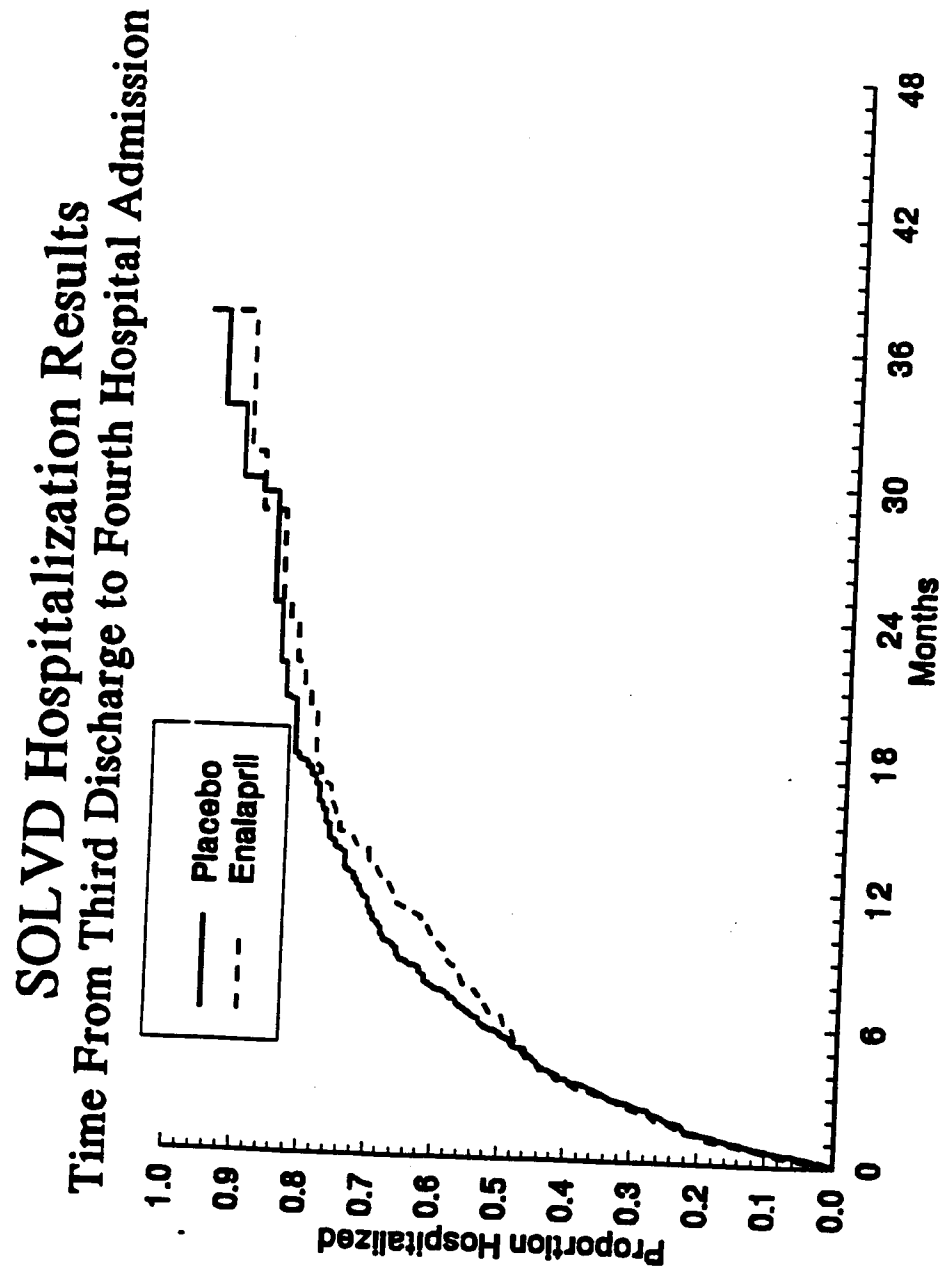

SOLVD Hospitalization Results
Time From Second Discharge to Third Hospital Admission
Placebo
Enalapril
Proportion Hospitalized
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
6
12
18
24
30
36
42
48
Months

Figure 9


SOLVD Hospitalization Results
Time From Third Discharge to Fourth Hospital Admission
Placebo
Enalapril
Proportion Hospitalized
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
6
12
18
24
30
36
42
48
Months

The implication here is that the interpretation of the covariate effects, or the difference between the log hazard ratios, may depend on the question of interest. For example, for the second event in the above example two questions are possible:

a) Will initializing enalapril therapy delay the occurrence of two hospital admissions?

or)

b) Given that a patient has had a first hospitalization, does enalapril delay a subsequent hospital admission?

With an unrestricted risk set and total event times, a positive treatment effect that is observed for a first event may appear to be sustained over subsequent events, even when no additional benefit after the first event exits.

Since methods with poor small sample properties may result in estimates that are either biased or highly variable, the parameter estimates of the various methods should be compared; however, since these methods differ with respect to the parameter being estimated, comparisons would be difficult to interpret. Parameter estimates from each method are estimated based on Cox's likelihood and are therefore asymptotically unbiased. Asymptotic properties of $\hat{\beta}$ from the Cox model are usually found to hold up under moderate and even small sample sizes except in situations with extreme censoring (Kalbfleisch and Prentice 1980, p.76). From these results, the parameter estimates were assumed to be unbiased; however, no

comparisons among methods were made since parameters from the different models have different meanings.

e. Constraining Covariate Effects

The models also differ with respect to whether covariate effects are estimated separately for each event or combined in some way over all events.

It is important to be able to estimate the effect of an important covariate, such as treatment, separately for each of the k failures. Covariate effects could be in different directions for the k events, favoring one treatment for one event and another treatment for another event. This could be a problem for any method and necessitates the ability to look at each event individually and test for homogeneity before combining the covariate effects into an overall summary statistic. Even if the direction of covariate effects differs over events, it may still be meaningful to combine these estimates for an "average" effect over all events.

The PWP, WLW and PC methods allow covariates to vary among events, while the AG method constrains the treatment effect to a single estimate. Although it is important to be able to estimate the main covariate separately, it may be desirable to have the ability to constrain other secondary covariates such as age or study center to a single estimate. We would like to control for these effects, but not use up degrees of freedom providing separate estimates for each event. This could result in lower power to detect treatment differences or exacerbate the problem of small sample sizes. The PWP method allows estimation of some covariates (*e.g.*, treatment

group) for each event, but also has the ability to constrain other covariates whose effects we would like to adjust for, but not necessarily estimate at each event. In the WLW and PC, all covariates are estimated separately for each event, which may result in a large number of parameter estimates.

f. Baseline Hazard/Proportional Hazard Assumptions

Another important model consideration is whether the baseline hazard is allowed to vary over the events or whether it is constant throughout the failure experience. Although the baseline hazard remains arbitrary or unspecified, PWP, WLW and PC methods all allow the hazard to change with each event. AG constrains the baseline hazard to be the same for all events.

All six methods assume proportional hazards. WLW, PWP and PC, which allow the baseline hazard to vary across events, also allow covariate effects to vary over events. The AG model constrains the baseline hazard to be the same for all events, and thus the covariate effects must be constant over all events. This is an important assumption in that it may make sense medically to assume that the hazard for an event and the impact of a covariate on the hazard change relative to event history.

g. Small Sample Results

The results of all six approaches are asymptotic; thus, inference based on these models is appropriate for large N and may not be valid for small or medium sized studies. However, since the choice of risk set, definition of event time and assumptions regarding

dependencies among event times differ among the six methods their small sample properties may differ. Therefore, the properties of the models in small sample situations should be examined.

ii. Limits of Investigation

There are some features of recurrent failure time data that might compromise the validity of any type of analysis. Since the models chosen cannot adjust for these limitations, the results of analyses based on these methods should be interpreted with caution if the following features are thought to be present.

a. Nonindependent Censoring

All 6 approaches require the assumptions of independent and noninformative censoring. If there is censoring that may affect the occurrence of the event of interest or the censoring is informative with respect to the regression parameters, then none of the models is valid.

One method that attempts to address the problem of nonindependent censoring events that are in the form of competing risks is the approach of Pepe (1991). As discussed in section II.C.i., her method acknowledges this type of censoring event and models the hazard of primary failure events conditional on whether the competing risk event has occurred. Pepe's method is directed at distinct events of different types, rather than recurrent events, and does not permit the inclusion of more than one covariate. Thus it is not appropriate for our application.

If nonindependent or informative censoring is suspected, it must be acknowledged so that parameter estimates and inference from the model can be interpreted cautiously.

b. Risk-Free Period

For some types of failure-time data, a period of time may lapse after an event before the patient is at risk for a subsequent event. An example where this might occur is the hospitalization after birth examples studied by AG. During the time that a patient is in the hospital, she is not at risk for another hospitalization and time to next hospitalization is therefore related to the length of hospital stay. Another example is the occurrence of MIs. There is a period of time after a patient suffers an MI when a subsequent event may not be considered a second distinct event but rather an extension of the previous event first event.

It is possible to account for this problem, at least in situations similar to the hospital admissions example, simply by redefining gap time. Instead of calculating gap time as the amount of time between the occurrence of events (k-1) and k, let gap time represent the length of time at risk for event k. Thus, the gap time for the k^{th} hospital admission, would be time from discharge after the $(k-1)^{th}$ admission until the k^{th} admission. This approach would not be helpful in the MI example unless there was a clearly defined rule specifying when an MI is considered a new event.

Chapter III

EVALUATION OF METHODS THROUGH SIMULATIONS

Simulations were used to compare the statistical properties of AG, WLW, PWP_2, PWP_3, PC and PC_m in an attempt to answer the following questions:

1) How does the degree of correlation among endpoints affect the performance of the tests? In particular, do tests based on models that require specification of the **dependence** structure (which by default assume recurrent event times are independent) have appropriate significance levels under the null hypothesis when assumptions are violated; *i.e.*, when the event times are correlated?

2) What is the impact of the choice of the risk set on the size and power of the tests? Although the meaning of the parameter estimates may be questionable for methods that used unrestricted risk sets, methods with more restrictive risk sets may have larger standard errors and, thus, poor power.

3) Since methods that model total time introduce correlation into the model, what is the effect on the size of the tests of modelling total time rather than gap time, when no true correlation between gap times exists?

4) Since all methods rely on asymptotic normality for inference, how are the test sizes affected by varying sample sizes? In addition to the number of subjects in the study, small sample properties may also be affected by the number of covariates in the model, number of events and degree of

censoring.

5) How do the methods perform relative to one another with respect to the power of the test to detect true treatment differences in the log hazard ratio?

6) Is there an increase in power by considering multiple events as opposed to an analysis of time to first event?

A. Simulation Parameters

The parameters of this simulation study were the following:

1) Survival distributions (including the number of failures, their failure time distribution and the degree and structure of the correlations among endpoints)

2) The degree of censoring

3) Sample size

4) Number and distribution of covariates

5) The value of β, which determines whether the simulation results are evaluating the procedure's size or power

The choice of the values of the simulation parameters is determined by which aspects of the models we want to evaluate and the specific questions we want to answer.

For all evaluations, data were generated using an exponential distribution. (Since all models evaluated here do not make parametric assumptions regarding this distribution, similar results would be expected with other survival distributions.) All event times were generated as gap times relative to the time of the previous event, since this is expected to be the case in applications of most interest here. For each patient the expected gap time from one event to the next is stationary. A weight variable, w, determined whether event times were correlated within patients or independent. This variable, which is also referred to as the degree of dependence, could range from 0 to 1. With the weight variable equal to 0, the expected value of the exponentially distributed gap times is the same for all subjects but as w increases, the difference in the expected value of event times among subjects increases. This increase in the between patient variation introduces within patient correlation among the gap times. Details of these calculations are given below.

Let

$$B_i=\exp(\beta' z_i)$$

be the i^{th} patient's factor to account for covariate effects. The patient is assigned a uniform random number, u_i, and the expected or mean gap time for that patient's events is

$$m_i=B_i[5\{(u_i-0.5)2w+1\}].$$

Again, each patient is assigned a uniform random number, this time for each of the k events, v_{ik}, and the patient's gap times are assigned as follows:

$$x_{ik} = -\ln(v_{ik}) B_i m_i.$$

The negative of the natural log of a standard uniform random number results in an exponential random number with an expected value of one.

Appendix I, Table 1 summarizes the actual correlations associated with the different values of the weight variable. For gap times, the correlation for a weight variable w = 0 was 0, while for w = 0.5 the correlation was 0.07 and for w = 1 the associated correlation was 0.22. Interestingly, the correlations for total event times were much higher. Even when gap times were independent, *i.e.* w = 0, the correlation between total times for events 1 and 2 was 0.69. These correlations were obtained empirically using random data.

A percentage of patients were randomly selected to be censored. For individuals selected to be censored, the censoring time was uniformly distributed between zero and the individual's total event time. Let u_i be a uniform random number assigned to the i^{th} patient. The censoring time, C_i, for the i^{th} individual selected to be censored is then defined as

$$C_i = u_i \sum_{k=1}^{s} x_{ik},$$

where x_{ik} are the gap times for patient i, and s is the total number of events.

For the covariate representing the treatment effect, values of zero and one were each assigned to half the individuals. Additional covariates, other than the dichotomous treatment variable, were generated as standard normal.

The effects of the simulation parameters on the performance of the various methods were examined one at a time by modifying that parameter and comparing the results to those from a standard set of parameter values. These standard values were: three recurrent events with no censoring, treatment group as the only covariate and a total sample size of 100. Results for this set of standard parameters are highlighted in bold in each table of simulation results. All simulations were run under conditions of both independent and correlated event times.

The number of simulations that were run depended on whether I was examining significance levels or power. Since the critical characteristic of a method is the true type I error rate, 10,000 simulations were run to obtain actual significance levels, while to investigate power, 1,000 were used. The standard error for the type I error rate was approximately 0.002, while for power the margin of error was approximately 0.007.

Table 2 summarizes the set of variables used for the simulations that examined the size of the tests and Table 3 summarizes the parameters used for simulations evaluating power. The corresponding table that summarizes the results from each simulation is also indicated.

The variables used in the simulations were chosen to answer the questions outlined in the beginning of the chapter. The values of the individual variables used in the simulations were chosen to demonstrate the impact of that variable on the true significance level or power of the individual tests.

TABLE 2

SUMMARY OF SIMULATION PARAMETERS

SIZE OF TEST

Degree of Depend.	Number of Events	Sample Size	Concom Variables	Percent Censored	True Beta	Results Table
0, .5, 1	3	100	1	0	0	4-8
0	1,2,4,5	100	1	0	0	5
1	2,4,5	100	1	0	0	5
0	3	20,500	1	0	0	6
1	3	20,500	1	0	0	6
0	3	100	3,5	0	0	7
1	3	100	3,5	0	0	7
0	3	100	1	50, 75	0	8
1	3	100	1	50, 75	0	8

TABLE 3

SUMMARY OF SIMULATION PARAMETERS

POWER

Degree of Depend.	Number of Events	Sample Size	Concom Variables	Percent Censored	True Beta	Results Table
0	3	100	1	0	0.20	9,11
0	3	100	1	0	0.40	9,10,11
0	3	100	1	0	0.60	9,11
1	3	100	1	0	0.20	9,11
1	3	100	1	0	0.40	9,10,11
1	3	100	1	0	0.60	9,11
0	3	100	1	50, 75	0.40	10
1	3	100	1	50, 75	0.40	10
0	3	100	1	0	0.10	11
0	3	100	1	0	0.30	11

To evaluate the methods, a summary statistic combining information over all events was used. The two PWP methods and Andersen and Gill both provided a straightforward Wald test using information over all events. WLW combines the individual parameter estimates and variances over all events using a weighted average where the weights are functions of the inverse of the covariance matrix. Details of this approach are shown is equations (2), (3) and (4), Chapter II. The WLW weighted average effects approach was also used to combine the event specific estimators of the Pepe and Cai and modified Pepe and Cai methods.

B. Simulation Results

i. Correlation

The first set of simulations is summarized in Table 4 and shows the effect of correlated event times on the size of a global test comparing two treatment groups with respect to time to a total of three events. The total sample size was 100 and there were no additional covariates other than the dichotomous treatment factor and no censoring. The purpose of this set of simulations was to compare methods that require specification of the dependence structure and either introduce correlation without adjusting for it or that don't account for correlation if it exists with the methods that adjust for correlation between event times. The sizes of tests from the six methods were compared using data generated with different degrees of correlation. With dependency of 0, the event gap time for all patients overall events was stable, with a mean of 5. For dependency of 50%, an individual's expected gap times remained stable over events but the expected values of event gap times among subjects varied from 2.5 to 7.5. For dependency = 1 the expected values of the event gap times varied from 0 to 10 over subjects, remaining stable over

the events for each subject.

PWP_2 is the Markov model that actually creates correlations between event times by modelling total times, but by default assumes the event times are independent. The true significance level for this method was clearly too high even when gap times were independent and the significance level increased as gap time correlation increased. With no correlation, the size of the test for this method was 0.209 and rose to 0.226 with dependency equal to 1.

When event times were independent, the size of the test for PWP_3, which models gap times, was 0.054, or very near the nominal significance level. This small inflation of the size of the test was probably real and due to the relatively small sample size. The size of the test did increase as the dependence between event times was introduced, and when the dependency parameter was set to one the size of the test was 0.108, which is clearly unacceptable.

The AG model did not do as poorly as PWP_2 in spite of the fact that it also models total time, and actually performed slightly better than PWP_3. The size of the test was 0.055 under independence and 0.090 when the dependency parameter was set to one. Thus although AG models total event times and uses an unrestricted risk set, the AG model which sums the likelihood overall events does not appear to be as sensitive to misspecification of the covariance structure as the method based on a stratified Cox model (PWP_2).

The following three methods attempt to adjust for the correlation that can be introduced (total times), real (correlated gap times) or both. The size of the test for WLW when event gap times were independent was

0.057, which was slighter higher than the corresponding significance levels for PWP_3 (0.054) and AG (0.055). This result suggests that this method may be more sensitive than PWP_3 and AG to small sample sizes. However the test is clearly unaffected by the amount of correlation that is introduced between gap times, suggesting that the method is correctly adjusting for the true correlation between gap times as well as for correlation that results from the method's approach to modelling totals times as opposed to gap times.

The Pepe and Cai method seems to have some small sample problems. The true significance level was 0.086, exceeding the nominal level even when there was no dependence between gap times. This result is probably due to the extra restriction placed on their risk set, causing it to become extremely small at later events. Although the increase in the significance level with increasing correlation is not dramatic, it does appear that this problem worsens in the presence of correlated event times. Perhaps correlated event times have the effect of further restricting the risk set, thus exacerbating the small sample problem.

The modified Pepe/Cai method models gap times instead of total times and therefore doesn't require the extra restriction on the risk set. The significance levels for this method were 0.063 with no correlation and 0.062 for dependency equal to one. It is obvious that although this method appears to have small sample problems, it is correctly adjusting for the correlation. The fact that the covariance estimator is the same for Pepe and Cai and the modified Pepe and Cai supports the hypothesis that the increase in the significance level for the PC method as the dependency increased is due to the effect of the dependency on the size of the risk set at later events, and not on failure of the method to adjust for correlation.

Table 4

Simulations - Effect of **Correlated** Event Times

Sample size = 100, Events = 3, All $\beta = 0$, One Covariate

Degree of Dependence	Size/Power					
	PWP_2	PWP_3	AG	WLW	PC	PC_m
0.00	**.209**	**.054**	**.055**	**.057**	**.086**	**.063**
0.50	.217	.068	.059	.056	.088	.067
1.00	**.226**	**.108**	**.090**	**.057**	**.100**	**.062**

ii. Number of Failure Events

The second set of simulations, summarized in Table 5, evaluates the effect of increasing the number of events on the type I error rate. The results were obtained both for conditions of no dependence and for the situation where dependence = 1; *i.e.*, where the expected value of an event varied over subjects. The number of events evaluated ranged from 1 to 5; however, correlation is not applicable to the situation with a single event. The sample size used for these results was 100 and there was no censoring.

PWP_2 resulted in an acceptable significance level only under the condition of one event: Its true significance level was 0.051. But even under independent conditions and only two events the significance level was 0.136 and for five failure events the significance level exceeded 0.30. Under the condition of dependency between gap times, the actual significance levels were even more inflated.

Both PWP_3 and AG did quite well under the independence assumption, with significance levels that were near the nominal levels regardless of number of events. When dependence between gap times was introduced, the type I error rates for both models increased as the number of events increased, although not nearly to the degree seen with the PWP_2 model. With five dependent failure events, the size the test for PWP_3 was 0.165, while for AG it was 0.172.

WLW did well with one or two events, but its significance level

did appear to rise slightly with 3 or more events. This increase in the size of the test associated with an increase in the number of events may be the result of small sample size problems, since the method estimates a vector of covariates for each event. In spite of the slight increase in size associated with increasing event numbers (its size never exceeded 0.060), it is important to note that the test is virtually unaffected by dependency. The results were the same whether or not event times were correlated.

The size of the test for the PC method was 0.051 with one failure event, but jumped to 0.062 for two independent events and exceeded 0.10 for five independent events. When the events were dependent, the size of the test was slightly but consistently higher, suggesting that the small sample problems due to the risk set definition may be exacerbated by dependence between event times.

Results for PC_m were similar to WLW except that a small increase occurred in the significance level for two events relative to one event, whereas this increase didn't appear for WLW until there were three events. After the increase in the test size from one event (0.051) to two independent events (0.061) or two dependent events (0.059), the significance levels remained consistent over event number. As with WLW, the increase is probably due to small sample problems.

Table 5

Simulations - Effect of Varying the **Number of events**

Sample size = 100, All β = 0, No censoring, One Covariate

Events	PWP_2	PWP_3	AG	WLW	PC	PC_m
1	.051	.051	.051	.054	.051	.051

No Dependence

Size/Power

Events	PWP_2	PWP_3	AG	WLW	PC	PC_m
2	.136	.055	.050	.053	.064	.061
3	**.209**	**.054**	**.055**	**.057**	**.086**	**.063**
4	.257	.049	.050	.058	.096	.063
5	.304	.046	.048	.060	.102	.058

Dependence

Size/Power

Events	PWP_2	PWP_3	AG	WLW	PC	PC_m
2	.142	.077	.068	.054	.079	.059
3	**.226**	**.108**	**.090**	**.057**	**.100**	**.062**
4	.287	.141	.115	.059	.110	.061
5	.341	.165	.172	.059	.119	.063

iii. Sample Size

Table 6 summarizes the relationship of sample size and type I error rates for dependent and independent event times. Each simulation included 3 events with no censoring. Sample sizes of 20, 100 and 500 were examined. Although survival trials are often much larger, if the asymptotic properties can be shown to hold for samples of 100 or 500, the results will extend to larger trials.

As has been shown in the previous simulation results the size of the test for PWP_2 always exceeds nominal levels; therefore, there is little benefit in the fact that the size of test is virtually unaffected by sample size.

Under independence the size of the tests for both PWP_3 and AG maintain near nominal levels regardless of the sample sizes within the range studied here. Both tests performed as well with only 20 observations, or 10 observations per treatment group, as they did with $n = 500$. When the times between events are dependent, the size of the test for these methods was too large, with no meaningful improvement as the sample size increased.

WLW's problem with small sample sizes was evident with a sample size as small as 20, where the true significance level was 0.075 under the condition of independent gap times. The method did show improvement when the sample size was increased from 20 to 100; with the size of the test dropping to 0.057, but even with 500 subjects the actual significance level remained slightly inflated at 0.055. The size of the test for WLW was unaffected by correlated event times regardless of the sample size. For most applications where survival analysis is used, sample sizes will be much larger than 20 or even 100 and WLW

asymptotic results would most likely continue to improve as sample size continues to increase. Samples as large 1000 are not uncommon in survival studies; however, the evaluation of sample sizes over 500 was out of the scope of this research due to the amount of computation required.

The PC method had even more problems with small samples than the WLW method and again, it is most likely due to their risk set. Under independence the size of the PC test ranged from 0.157 for a sample size of 20 to 0.064 for a sample size of 500. The corresponding true significance levels under dependent gap times ranged from 0.175 to 0.082. The PC method did appear to be approaching the nominal significance levels, but very slowly, and significance levels appeared to be increased when the gap times were correlated.

Modified Pepe and Cai had small sample results that were slightly worse than WLW for a sample size of 20, but there was a definite tendency for the true significance levels to decrease as the sample size increased. For a sample size of 500, the difference between the size of the modified PC and WLW was within the margin of error of the simulations. In addition, the size of test for PC_m was unaffected by dependent gap times, suggesting that the method is controlling for the correlation as well as WLW.

Table 6

Simulations Investigating Small Sample Properties

Effect of Varying **Sample Size**

Events = 3, All β = 0, No censoring, One covariate

No Dependence

Sample Size	**Size/Power**					
	PWP_2	PWP_3	AG	WLW	PC	PC_m
20	.205	.051	.055	.075	.157	.080
100	**.209**	**.054**	**.055**	**.057**	**.086**	**.063**
500	.214	.051	.052	.055	.064	.056

Dependence

Sample Size	**Size/Power**					
	PWP_2	PWP_3	AG	WLW	PC	PC_m
20	.233	.119	.100	.065	.173	.079
100	**.226**	**.108**	**.090**	**.057**	**.100**	**.062**
500	.224	.112	.091	.054	.082	.058

iv. Number of Covariates

The next factor that was investigated through simulations was the effect of increasing the number of covariates on size of the tests. These results are summarized in Table 7. The single dichotomous "treatment" effect was always included, and the effect of adding 2 or 4 additional covariates was examined. All covariates in addition to treatment were generated as standard normal and all betas were equal to zero. As with all other previous simulations, the size of the test that was evaluated was for the null hypothesis of no difference between 2 treatments in time to an event with no censoring and a the sample size of 100. Evaluation of the tests was made for independent and correlated gap times.

The size of the tests for PWP_2 and PWP_3 both increased slightly as more covariates were added to the model. As was always the case for these two models, the significance levels were consistently higher under the condition of correlated gap times, and the relative increases in the size of the test as the number of covariates increased were similar. The size of the test for AG did not increase with the number of covariates under independent gap time conditions, but rose from 0.090 to 0.10 when event times were correlated. The results for WLW and PC_m were similar. While the sizes of the tests for WLW and PC_m both increased slightly when additional covariates were included, the effects of additional covariates were similar under both dependent and independent event times. Interestingly, the size of the test for PC appeared to increase with the number of covariates under conditions of no dependence, but not when dependent gap times were generated.

Table 7

Simulations Investigating Small Sample Properties

Effect of Varying the Number of **Covariates**

Sample size = 100, Events = 3, All β = 0, No censoring

No Dependence

Concomitant Variables	**Size/Power** PWP_2	PWP_3	AG	WLW	PC	PC_m
1	**.209**	**.054**	**.055**	**.057**	**.086**	**.063**
3	.218	.057	.055	.069	.086	.067
5	.222	.059	.053	.072	.103	.077

Dependence

Concomitant Variables	**Size/Power** PWP_2	PWP_3	AG	WLW	PC	PC_m
1	**.226**	**.108**	**.090**	**.057**	**.100**	**.062**
3	.227	.117	.094	.068	.096	.072
5	.235	.123	.100	.072	.099	.074

v. Censoring

The effect of censoring on the true significance levels of the tests was also investigated. The percent of patients censored ranged from 0 to 75%. These results are presented in Table 8. Except for WLW the size of the tests for all methods dropped when censoring was introduced and this decrease in the true significance level was most notable for the two Pepe and Cai based methods. For independent event times the true significance level for the PC method dropped from 0.086 with no censoring to 0.026 when 75% of the subjects were censored. The corresponding change in the significance level for the PC_m was from 0.063 with no censoring to a conservative 0.012 with 75% censoring. For all methods, the relative decline in the true significance level as the censoring increased was the same regardless of whether event times were correlated. Thus, WLW seems be the only method with significance levels that are unaffected by censoring. The decline in significance levels that was seen for the other methods cannot be interpreted as an improvement in the statistical properties of the methods due to the introduction of censoring, but a different mechanism that is influencing the size of the test in the opposite direction. Furthermore, if the effect of censoring on the data generated is to reduce the sample size, then results are inconsistent with the results for other variables that were suspected of creating small sample problems: Decrease in sample size, increase in number of covariates and increase in number of events, all of which were associated with elevated significance levels. This may be because for all methods except WLW censoring results in less information and reduced power, (*i.e.*, reduced risk sets for later events) while WLW uses information from the first event in obtaining parameter estimates for all events.

Table 8

Simulations

Effect of Varying the Degree of **Censoring**

Sample size = 100, Events = 3, All β = 0, One Covariate

No Dependence

Percent	**Size/Power**					
Censored	PWP_2	PWP_3	AG	WLW	PC	PC_m
0	**.209**	**.054**	**.055**	**.057**	**.086**	**.063**
50	.163	.046	.041	.056	.041	.018
75	.134	.042	.029	.057	.026	.012

Dependence

Percent	**Size/Power**					
Censored	PWP_2	PWP_3	AG	WLW	PC	PC_m
0	**.226**	**.108**	**.090**	**.057**	**.100**	**.062**
50	.179	.089	.077	.055	.044	.023
75	.140	.072	.058	.058	.028	.016

vii. Power - Comparison of Multiple Event Methods

Although it's understood that the tests with inflated significance levels are not valid, the power results for all six methods included in the preceding simulations were evaluated and are presented in Tables 9 and 10. Based on the size of the test, PWP_2 should never be used and PWP_3 and AG should not be used when event times are dependent; power results for these methods under the corresponding conditions are not discussed. The true significance level of PC was also larger than nominal levels under most conditions. Power for this method is also not discussed in detail.

Power was investigated under several conditions. In Table 9, the β varied from 0.2 to 0.6. The sample size was set to 100, there were 3 events and no censoring. The parameter β, or log hazard ratio for the treatment effect, was the same for each of the three events. A log hazard ratio of 0.2 represents a 22% increase in hazard for treatment group 1 relative to treatment group 0, while a log hazard ratio of 0.6 represents an 82% increase.

All of the valid methods had adequate power to detect $\beta=0.6$ both when events were dependent and when they were independent. For $\beta=0.2$, however, none of the valid methods had more than about 40% power.

For $\beta=0.4$, all methods had greater than 85% power under independence, except for PC which resulted in only 60% power. The power was greater for PWP_3 (93%), AG (94%) and PC_m (94%) than for WLW (86%). When event times were dependent, the power for all methods was reduced but the only methods that remained valid in the presence of correlated event times were WLW and PC_m. The power from

the PC_m method was slightly greater than the power for WLW, 60% vs. 53% under dependence with $\beta = 0.4$.

Table 9

Simulations - Power with Different β

Sample size = 100, Events = 3, No Censoring

No Dependence

	Size/Power					
True Beta	PWP_2	PWP_3	AG	WLW	PC	PC_m
0.20	**.578**	**.382**	**.393**	**.310**	**.199**	**.413**
0.40	.971	.931	.941	.857	.600	.937
0.60	.999	.999	.999	.999	.885	.999

Dependence

	Size/Power					
True Beta	PWP_2	PWP_3	AG	WLW	PC	PC_m
0.20	**.426**	**.293**	**.269**	**.192**	**.191**	**.207**
0.40	.795	.723	.682	.531	.388	.602
0.60	.966	.954	.945	.854	.705	.914

The second set of power simulations are summarized in Table 10 and the results illustrate the impact of different degrees of censoring on each method in their ability to detect $\beta = 0.4$. These results were run under conditions of both dependent and independent event times. The sample size was 100 and there were three events.

As expected censoring had the effect of reducing the power for all methods. Reductions in power were greater for the PC_m procedure than for WLW, PWP_3 or AG. Under independence, the power from the PC_m model fell from 94% with no censoring to 33% when 75% of the observations were censored. The corresponding drop in power for WLW was from 86% with no censoring to 54% for 75% censoring. The relative reductions in power for PWP_3 and AG were similar to WLW. PWP_2 and PC were not valid tests.

With correlated gap times the power for PC_m fell from 60% (0 Censoring) to 16% (75% censoring), while the drop in power for WLW was less extreme (53% to 31%). The PWP_3 and AG models were not valid when event times were correlated.

Table 10

Simulations - Power: Effect of Varying degrees of **Censoring**

Sample size = 100, Events = 3, $\beta = 0.4$

No Dependence

Percent	**Size/Power**					
Censored	PWP_2	PWP_3	AG	WLW	PC	PC_m
0	.971	.931	.941	.857	.600	.937
50	.874	.770	.771	.665	.354	.616
75	.787	.657	.612	.536	.273	.329

Dependence

Percent	**Size/Power**					
Censored	PWP_2	PWP_3	AG	WLW	PC	PC_m
0	**.795**	**.723**	**.682**	**.531**	**.388**	**.602**
50	.654	.546	.517	.373	.223	.295
75	.552	.438	.386	.309	.158	.163

To summarize the power analysis, under conditions of independent event times and no censoring PWP_3, AG and PC_m were more powerful than WLW. Based on significance levels from the simulation studies, neither PWP_2 or PC were valid and so power comparisons were inappropriate. When event times were correlated PC_m was still slightly more powerful than WLW but only in the absence of censoring. (Note that PWP_3 and AG were not valid methods when gap times were correlated due to inflated significance levels.)

When censoring was introduced and gap times were independent, WLW was more powerful than PC_m but not as powerful as PWP_3 and AG; however as shown in Table 8, the significance levels of PC_m, AG and PWP_3 are questionable with censored data. WLW remained more powerful than PC_m in the presence of censoring when correlated gap times were introduced.

viii. Power - Single vs. Multiple Events Analysis

The last set of simulations compared the power of an analysis that uses information from multiple events with Cox's analysis of time-to-first event. These simulations were designed to investigate whether it is worthwhile to use a method that considers all events relative to a time-to-first event analysis in terms of improving the power to detect covariate effects. The WLW method which is based on all events was chosen as the multiple events analysis, since it was the only method that was valid in terms of adequate significance levels under conditions of correlated event times as well as censoring. The simulations compared the two methods under different sizes and patterns of covariate effects over events to identify situations where the multiple events analysis may be beneficial in terms of power, as opposed to situations where it might actually be less powerful.

The simulations were run under conditions of 4 events with 100 patients, no censoring and no dependence between event times. The parameter β ranged from 0 to 0.6 and the pattern of the covariate effects varied over the 4 events. When $\beta=0$ over all events, the true significance levels of the two methods were equivalent. When a treatment effect was present at the first event only, the Cox model was more powerful than WLW, regardless of the size of the effect. In this simulated study of 100 patients, power for the Cox model wasn't adequate unless there was a relatively large treatment effect, $\beta=0.6$ (power=0.84) corresponding to a relative risk of 1.82. The corresponding power for WLW with $\beta=0.6$ at the first event was 0.68.

The point where the power for the two methods was approximately the same occurred when half the effect from the first event was present at

the second event (*e.g.*, $\beta = 0.4$ for event 1 and $\beta = 0.2$ for event 2). Once a treatment effect, or β, of equal magnitude was present in the first 2 events, WLW was more powerful than the Cox first event analysis. The advantage of WLW increased dramatically as the effect persisted over more events. WLW provided adequate power to detect $\beta = 0.4$ when the effect was present over the first 3 events (power = 0.83) while for the Cox model the power to detect $\beta = 0.4$ was only 0.52. When the treatment effect of 0.4 persisted over all 4 events the difference in power was even greater, increased from 0.51 for the first event analysis to 0.91 for the WLW.

It is important to note that since this power analysis was run under conditions of independent event times, the benefits shown for these multiple event analyses relative to a time to first event were probably greater than if event times were correlated. With highly correlated gap times, not much additional independent information can be obtained from multiple events and power based on a multiple time-to-event analysis is reduced. This result was demonstrated in the simulation studies shown in Tables 9 and 10. In addition, results in Table 11 correspond to simulated data without censoring. Censoring will reduce the power of both methods, but may have a greater impact on the power to detect a persistent treatment effect over events.

To summarize, a multiple events analysis may result in increased power relative to a first event analysis when more than half the treatment effect from the first event is present at a second event. Dramatic improvements can be expected when greater effects persist over more events. The actual gain in power will depend on the amount of correlation among events and may also depend on the amount of censoring.

Table 11

Power for WLW vs. Cox's Model

Events = 4, N = 100, Treatments = 2, Dependence = 0

True β for Event Number				Size/Power	
1	2	3	4	WLW	Cox
0.0	0.0	0.0	0.0	0.057	0.056
0.2	0.0	0.0	0.0	0.163	0.173
0.4	0.0	0.0	0.0	0.374	0.496
0.6	0.0	0.0	0.0	0.676	0.837
0.2	0.2	0.0	0.0	0.239	0.174
0.4	0.4	0.0	0.0	0.665	0.515
0.6	0.6	0.0	0.0	0.923	0.833
0.2	0.2	0.2	0.0	0.313	0.180
0.4	0.4	0.4	0.0	0.831	0.520
0.6	0.6	0.6	0.0	0.986	0.833
0.2	0.2	0.2	0.2	0.389	0.157
0.4	0.4	0.4	0.4	0.913	0.514
0.6	0.6	0.6	0.6	0.994	0.838
0.2	0.1	0.0	0.0	0.200	0.183
0.4	0.2	0.0	0.0	0.516	0.483
0.6	0.3	0.0	0.0	0.848	0.839

C. Summary

The following section summarizes the simulation results in terms of the questions outlined at the beginning of the chapter.

1) The first question the simulations addressed was the effect of correlation among endpoints on the different tests. The simulation results indicate that the size of the true significance level for PWP_2, which introduces correlation through the model, is always inflated. For PWP_3 and AG, the sizes were adequate under most conditions unless there was correlation among endpoints. When gap times were independent and there was no censoring the asymptotic results for these two methods were excellent. WLW and PC_m both adjusted for correlation and performed better than any of the other methods when event times were correlated. Therefore, as expected, when event times were correlated methods that assume independence did not perform well (and may give misleading results).

2)The different risk sets appeared to have more of an impact on the interpretation of model parameters than on the performance of the tests, except in the case of PC, where the restrictions placed on the risk set appear to affect the significance levels at later events. The restricted risk sets of PWP_2, PWP_3 and PC_m are the same and depend on event number but not event time. Since PC_m performs well under most conditions and PWP_3 performs well under independence, when the test is valid, this type of risk set does not adversely affect the performance of the tests. AG and WLW both use unrestricted risk sets, and although this makes interpretation of parameters difficult, the statistical properties of the tests were as expected.

3) The next issue was related to the performance of methods that introduce correlation by modelling total time rather than gap time. PWP_2, AG, PC and WLW methods all model total times. WLW and PC adjust for correlations through the covariance structure. As noted above, the significance level of PWP_2 is always inflated, while PWP_3 is inflated only when correlated gap times are introduced. Since the only difference between the two methods is that PWP_3 models gap times instead of total times, the impact of modelling total times with a stratified Cox model is a dramatic increase in the actual significance level. Surprisingly, AG, which also models total times without adjusting for correlation, performs nearly as well as PWP_3. The properties of AG were excellent except when true correlation among gap times was introduced. The difference is that PWP_2, the stratified Cox model, calculates the likelihood within each stratum and then multiplies them across strata, whereas AG models the event rate, calculating the likelihood over all events. The stratified Cox model is extremely sensitive to the assumption of independent strata. The event rate model of AG is less sensitive to correlations introduced through the use of total times.

4) As noted above, the small sample properties of PWP_3 and AG were excellent when event times were uncorrelated. WLW and PC_m did not have as good asymptotic results as PWP_3 and AG when event times were independent, but performed much better when endpoints were correlated. The significance level of PC was always above the nominal level and the method seemed sensitive to small samples and conditions that tended to reduce effective sample size. The size of PWP_2 was always much too large, although this was probably not related to small sample properties.

Finally, WLW was the only method that performed well under censoring. For all other methods actual significance levels decreased dramatically as the amount of censoring increased.

5) The results of the power analysis focused on only those methods which were found to be valid in terms of actual significance levels. Under conditions of independent event times and no censoring PWP_3, AG and PC_m were more powerful than WLW. WLW and PC_m were the only methods that were valid under dependence. Under these conditions, PC_m was the more powerful when there was no censoring, while WLW was more powerful when censoring was present.

6) In terms of power relative to an analysis of time to first event, the multiple event approach was beneficial when at least 50% of the treatment effect was still available at the 2nd event. Dramatic improvements in power might be expected when effects persist over more events.

Based on these results the following recommendations are made:

In most cases WLW will perform well for evaluating covariate effects overall events. If event times are known to be independent and censoring is not extreme, PWP_3 or AG could be used for analyzing covariate effects overall events and may have better power than WLW. The individual event parameters from PWP_3 are the most appropriate for evaluating the effects of covariates at individual events.

Chapter IV
EXAMPLES

The following examples came from well-known double-blind, randomized, cardiovascular clinical trials. Although the primary outcome was mortality in each, there were several important secondary outcomes such as the occurrence of MI and hospital admissions which were recurrent events. The six methods for analyzing recurrent failure events that were evaluated through simulations in the preceding chapter were applied to these examples in order to compare and contrast them, and to see how the results differed from a traditional single-event survival analysis method. As seen in the simulation studies true significance levels for PC and PC_m decreased dramatically in the presence of censoring, which may be due to the way these methods were implemented. Since the censoring in the following examples was extreme, results from PC and PC_m were provided in the tables but were not discussed in detail.

A. Hospitalization in Severe CHF - CONSENSUS

i. Study Description

Congestive heart failure (CHF) is a relatively common disorder with a poor prognosis. One year mortality has been estimated to be as high as 50% in patients with the most severe form of CHF (CONSENSUS 1987) and approximately 35% of the patients diagnosed with CHF are hospitalized yearly (SOLVD 1991).

The CONSENSUS trial was a randomized, double-blind study that compared the effect on mortality of the angiotensin-converting-enzyme (ACE) inhibitor enalapril vs. placebo in patients with severe CHF (New York Heart Association [NYHA] functional class IV) taking conventional therapy (CONSENSUS 1987). For this class of patients annual mortality is estimated to be greater than 50%. It was thought that ACE inhibitors might reduce deaths and rates of hospitalization by improving ejection fraction and exercise capacity, decreasing signs and symptoms of congestion and also decreasing preload and afterload.

The planned sample size for the CONSENSUS study was 400 patients; however, approximately 20 months after the first patient was enrolled, the ethical review committee reviewed the data on 244 patients and found that the difference between the two treatment groups in mortality was probably not due to chance. Study enrollment and follow-up were stopped. At the time the study was stopped, 253 patients had been allocated to receive enalapril (n = 127) or placebo (n = 126). Follow-up ranged from 1 day to 20 months, with a mean of 188 days. Six month follow-up was complete on 194 patients and 102 patients had 12 months of follow-up.

ii. Results - Effect of Treatment

Table 12 summarizes the important baseline characteristics of the patients as well as their hospitalization history by treatment group. Ejection fractions were not measured in this study. The average age at baseline for the two groups was approximately 70 years. Slightly more than half the patients had a previous myocardial infarction and between one fifth and one fourth of the patients were diabetic. Seventy-five patients in the placebo group (60%) and 74 patients in the enalapril

group (58%) were hospitalized at least once during the trial.

In addition to the Cox's proportional hazards model for time to first hospital admission, AG, PWP_2, PWP_3, WLW, PC and PC_m were used to analyze time to hospitalization, including up to four hospital admissions per subject. In addition to the treatment effect, models from each method adjusted for the patient's age at baseline, a dichotomous variable indicating whether the patient had a previous myocardial infarction and a dichotomous variable indicating whether or not the patient was diabetic. The AG, PWP_2 and PWP_3 models provided a Wald Chi-square for treatment effect summarized over all events. Overall test statistics for WLW, PC and PC_m were obtained by using the WLW formula in Chapter II, equations 2, 3 and 4. The results are summarized in Table 13. Parameter estimates and standard errors are shown for each event, except for AG, as well as combined over all events. The relative risk ($\hat{RR}$) was calculated as $\exp\{\hat{\beta}\}$ and refers to the overall treatment effect (except for the Cox model where it refers to the first event). A relative risk less than one represents a decrease in risk of hospitalization for the enalapril group.

Cox's model resulted in an estimated treatment effect of -0.28 with a standard error (se) of 0.17, which corresponds to a multiplicative change in the hazard, or relative risk, of 76%. Thus the risk of a first hospitalization is estimated to be reduced by 24% for the enalapril group relative to the placebo group. The p-value from the Cox's model (0.102) indicates that the difference between treatments is not statistically significant.

The parameter estimate (se) for a treatment effect over all events from

the PWP_2 model was $\hat{\beta}$ =-0.33 (0.12). This estimate corresponds to a relative risk of 72%. The individual event parameter estimates (se) from PWP_2 suggest a moderate treatment effect favoring enalapril at the first event, $\hat{\beta}$ =-0.27 (0.17), corresponding to an $\hat{RR}$ =76%. The estimated treatment effect for the first event differed slightly from the Cox model because the effect of the three additional covariates, age, previous MI and diabetes, are averaged over all events. When the additional covariates were estimated separately for each event or excluded from the model, the first event parameter estimate was the same as the Cox model estimate. A stronger effect was observed at the second event ($\hat{RR}$ =53%), but by the third event the parameter estimate actually favored placebo, $\hat{RR}$ =116%, or a 16% increase in risk for the enalapril group. By the fourth event, the treatment effect was similar to the second event. Standard errors from the parameter estimates increased with event number.

For all methods that provide parameter estimates for individual events, the meaning of the parameter for time to the first hospital admission is the same. After the first event, the meaning of the parameters differ among the models. The individual parameters, or relative risk estimates from PWP_2 represent the change in risk of hospitalization for the enalapril group relative to the placebo group as measured from the start of the trial. For example, for the second event, there is a 47% reduction in risk of a second hospitalization in the enalapril group compared to the placebo group, as measured from the start of the trial. The interpretation of the individual model parameters form PWP_2 is questionable because the method models total time since study start. The length of time to a second admission will be long if the time to the first admission is long, even if the second admission occurs shortly after the first. Thus the

parameter estimates do not represent the beneficial effect of enalapril on prolonging time between hospital admissions and may be misleading.

The p-value for an overall treatment effect for the PWP_2 analysis was 0.007. This result should be interpreted cautiously since simulation studies showed the true significance level from this method was inflated. The method models total times and does not adjust for correlations, either introduced or real. In other words, inference from this model assumes that the off-diagonal elements of the covariance matrix for the parameter estimates are zero. The covariance matrix obtained from the WLW method, which also models total time but then provides estimated correlations between parameter estimates, is included in Appendix II, Table 1 and shows that the off-diagonal elements of the covariance matrix are not zero. So PWP_2 assumes that information contributed at each event time is independent and since correlations exist between events, the test is not valid and may be exaggerating the evidence of a treatment effect over all events.

PWP_3 differs from PWP_2 in that it models gap times instead of total times. The parameter estimate for the effect of treatment on time to hospital admission combined overall events from PWP_3 was similar to the PWP_2 model: $\hat{\beta} = -0.31$ (0.12) or $\hat{RR} = 73\%$. For the first event, the estimate was the same as the Cox model and similar to PWP_2, indicating a moderate effect favoring enalapril. By the second event the PWP_3 estimate indicated a greater treatment effect than for the first event, $\hat{\beta} = -0.54$ (0.24) or $\hat{RR} = 58\%$ but the magnitude of the effect was not as large as was seen with the PWP_2 model. The estimated parameter at the third event was similar to PWP_2, suggesting a small treatment effect in the opposite direction from the first two events; *i.e.*, favoring placebo.

The relative risk for the fourth hospitalization was 49% ($\hat{\beta}$ =-.71, se = .44) again favoring enalapril. This estimate represents a decrease in the relative risk (*i.e.*, increased treatment effect) from the PWP_2 model of 5 percentage points. As with PWP_2, standard errors of the estimates increased with increasing event number.

Because PWP_3 models gaps times and includes in the risk set only those patients who have had a previous event, parameter estimates for a treatment effect at the individual events from the PWP_3 model are probably the most meaningful among all the models. The event-specific estimates from the PWP_3 model represent the effect of treatment on the length of time between events, not since the start of the trial. In this example, the estimate for the second event represents a 42% reduction in the risk of hospitalization in patients who have had one admission, as measured from the time of the first admission. This is the most straightforward representation of the effects of therapy on developing a second event, after the occurrence of a first event, since the time to the second will not be artificially influenced by the length of time prior to the first event.

The p-value for an overall treatment effect based on the PWP_3 analysis was 0.012, which was greater than the p-value from the PWP_2 analysis. As with PWP_2 inference from this model should be interpreted cautiously. Although the model does not introduce correlations, the covariance matrix will not adjust for correlations if they exist. Simulation studies showed the true significance level from this method was inflated when gap times were correlated. Inference from this model assumes that the off-diagonal elements of the covariance matrix for the parameter estimates are zero. The covariance matrix obtained from PC_m which also

models gap times and has the same risk set, but estimates correlations between parameter estimates, is shown in Appendix II, Table 1. Although variance and covariance estimates at later events are inflated, correlation is evident between the first two events. So PWP_3 assumes that the information contributed at each event time is independent and since correlations exist between parameter estimates for individual events, the test is not valid and may be exaggerating the evidence of a treatment effect over all events. As expected, this effect does not appear to be as pronounced as for PWP_2.

Because the individual parameter estimates and standard errors from PWP_3 are meaningful and associated inferences do not require assumptions of independence, results from hypothesis tests about the individual events are valid. In this study the treatment effect at the first event is not quite significant, and the evidence for a treatment effect from the first to the second admission is stronger, with a p-value less than 0.05. There is no evidence of a treatment effect for time from the second event to the third event, where the direction of the estimate actually switches, favoring the placebo group. By the fourth event the parameter estimate again favors the enalapril group, but the standard error is so large that the effect was nonsignificant.

The next method was AG, which provides parameter estimates and results over all events only. The estimated treatment effect over all hospital admissions was a relative risk of 76%, $\hat{\beta}$=-0.28 with a standard error of 0.12. This is slightly lower than the estimates from either of the PWP models.

AG models the overall hospitalization rate and the parameter estimate

represents the difference between treatments in risk of hospital admission, regardless of past hospitalization history. From this example the AG model indicates that among all patients, there is an overall reduction in risk of hospitalization of 24% for the enalapril group.

The p-value associated with the treatment effect from the AG model was $p = 0.023$. As with the PWP models, the AG model is probably not a valid model for these data, since there appears to be correlation among the parameter estimates. Significance levels were near the nominal level for AG when gap times were independent, but not in the presence of correlation in the simulations.

Based on the WLW model, the estimated treatment effect over all events was -0.30 (se = 0.16), representing a relative risk of 74%. The parameter estimate for the first event was the same as from the Cox model. For the second event, the WLW estimate $\hat{\beta} = -.55$ (se = 0.23) was similar to the PWP_3 estimate, corresponding to a relative risk of 58% or a reduction in risk of 42%. WLW differed from the previous models in the estimate of the treatment effect for the third event: an 84% relative risk still favoring the enalapril group ($\hat{\beta} = -0.17$, se = 0.32). The parameter estimate from WLW at the fourth event represented a relative risk of 64%, a smaller effect than was seen for the two PWP models. Standard errors for the parameter estimates of the individual events were similar to those from the PWP models.

WLW models total times and includes all patients in the risk set, in what they called a marginal approach. Because of this approach, interpretation of parameter estimates from individual events for the WLW model is difficult. A good example of the impact of the WLW risk set on the

interpretation of the parameters is found for the third event parameter estimates. This parameter represents a 16% reduction in risk of a third hospitalization for the enalapril group which is in the **opposite** direction of the effect estimated from PWP_3. This happened for two reasons: 1) WLW included all patients in the risk set, rather than just those patients with at least 2 hospitalizations. In this regard, note that when the third hospital admission rate was calculated among all patients (representing the WLW type risk set) it was only slightly lower in the placebo group (19/126 = 15%) than in the enalapril group (25/127 = 20%), while for patients with at least 2 hospitalizations (the PWP_3 risk set) the rate was much lower in the placebo group (19/39 = 48% vs. 25/35 = 71%). 2) WLW is based on time from randomization rather than time from second admission, and since first and second hospital admissions tended to occur later in the enalapril group than in the placebo group, effects of enalapril on the first two events carry over into WLW's estimate of treatment effect for the 3rd event. So in spite of a crude hospitalization rate still favoring placebo, the covariate estimate from WLW actually favors enalapril due to the influence of the first two events. However, the WLW estimate for an effect over all events from this model appears to be reasonable relative to the PWP_3 estimate.

In spite of the problems with individual event estimates, WLW seemed to have the best statistical properties based on simulations, particularly in the presence of censoring. The Wald chi-square from WLW was 3.39, $p = 0.065$. Although the p-value from WLW is lower than the p-value from the Cox model, it is still not quite significant. Therefore, after adjusting for the correlation among event times, there is not quite enough additional information from the second through fourth hospital admissions to declare the treatment effect statistically significant.

Results for PC and PC_m are shown in the table but are not summarized in detail. Although individual event and averaged parameter estimates for the treatment effects from PC_m were similar to those observed for PWP_3 and somewhat similar for PC, the two methods as implemented here had problems with increasing standard errors at the later events. This is probably due to a combination of the small sample size and high proportion of censored patients.

iv. Results - Additional Covariates

The effects of three additional covariates, age, whether the patient had a previous MI and whether the patient was diabetic, were also examined. Table 14 summarizes the analysis of the effect of age at baseline on hospital admissions controlling for treatment, history of MI and diabetes status. Although we might expect the risk of hospitalization to be higher in older patients, there was no evidence of an age effect in this study. This may be because age is confounded with MI history and diabetes status (*i.e.*, older patients are more likely to have had an MI and/or be diabetic) and both these factors were also included in the models.

The results for the effect of having a previous MI on risk of hospital admissions in models that include treatment, age and diabetes status are summarized in Table 15. The results from the Cox, PWP_2, PWP_3, AG and WLW models all indicated that having a previous MI significantly increased the risk of hospitalization (all $p < 0.01$). The parameter estimate from the WLW analysis overall events indicated patients with a previous MI had a 62% higher risk of hospital admissions than patients with no MI. Individual event estimates from the PWP_3 model suggested important increases in risk for the first and third events only.

Table 16 summarizes the analysis of the effect of having diabetes on the risk of hospital admissions in models that also included treatment, age and whether a patient had a previous MI. While there was no evidence of an effect of diabetes from the Cox model ($p = 0.290$), the models that consider evidence from all events suggested that having diabetes increased the risk of hospitalization. The results from the WLW model suggested a 39% increase the risk of hospital admission associated with having diabetes, however this effect was not quite significant, $p = 0.079$. The individual event parameter estimates from the PWP_3 model suggested a small nonsignificant increase in risk for the first event, with large significant relative risk estimates for the second and fourth events. In this case the results from the WLW analysis may be diluted with the small nonsignificant effect observed for the first admission. Although results from PWP_2, PWP_3 and AG were all highly significant, the validity of these methods is questionable.

iv. Summary

Even though the individual event parameters from the different models were estimating different things, the overall parameter estimates of the effect of treatment on hospital admissions were fairly consistent among the methods. However there was wide variation in the standard errors and associated p-values.

Conclusions from this analysis with respect to the effect of enalapril on risk of hospitalization should be based on inference from the WLW method. This analysis suggested that there was probably a reduction in the risk of hospitalization due to enalapril; however, since the p-value was not ≤ 0.05, definitive conclusions could not be made. Examination of the pattern of covariate effects over the events should be based on

the PWP_3 method, which suggests that enalapril tended to reduce the risk of most hospital admission events, although the difference was significant for the second event only.

The effects of three additional covariates were also examined in models that included all covariates, as well as treatment. Having a previous MI increased the risk of hospitalization in these patients. Being diabetic also appeared to increase the risk of hospitalization; however, the results from this analysis are inconclusive. In this study age was not associated with an increased risk of hospitalization. This last result is counter-intuitive and may be due to the confounding of age with diabetic status and MI history.

The original CONSENSUS study found a significant reduction in mortality in the enalapril group. At the end of the study the crude reduction in mortality due to enalapril was 27%, $p = 0.003$. Mortality in these trials was high and not only is it a competing risk for hospitalization, the outcome of interest, it occurred at a differential rate in the two treatment groups. Since the methods used here assume independent non-informative censoring, the results must be interpreted cautiously. In this example death will prevent the occurrence of the hospitalization endpoint and it seems reasonable to assume that patients who died would have had a high rate of hospital admissions if they had lived. Since death occurs more frequently in the placebo group, we may be seeing fewer hospitalizations in the placebo group than we would expect without a competing risk. If this is the case, then our estimate of a treatment effect with respect to hospital admissions may be conservative. If a significant or important treatment effect can be demonstrated in the presence of conservative bias against a drug, then the problem of informative, dependent censoring is less important. Bias in the other

direction, possibly enhancing a treatment effect, is much more troublesome and would threaten the validity of a trial.

Pepe's approach to incorporating different endpoints that might be competing risks could be considered for this example (Pepe 1991).

Table 12
Consensus Trial
Subject Characteristics and Hospitalization History

Variable	Placebo (n = 126)	Enalapril (n = 127)
Age (yr)		
mean	69.7	70.6
s.d.	8.5	9.8
Patient History		
Number (%) with:		
Previous MI	67 (53.2%)	65 (51.2%)
Diabetes	27 (21.4%)	29 (22.8%)
Patients Hospitalized		
Number (%) with:		
1 admission	75 (59.5%)	74 (58.3%)
2 admissions	39 (31.0%)	35 (27.6%)
3 admissions	19 (15.1%)	25 (19.7%)
4 admissions	10 (7.9%)	13 (10.2%)

Table 13
Consensus Trial
Analysis of Treatment Effect on Hospital Admissions
Adjusting for Age, History of MI and Diabetes Status

Model	Parameter Estimates (standard error) Admission Number: 1	2	3	4	All	$\hat{RR}$*	Chi-Square	P-value
Cox	-.28 (.17)	--	--	--	--	.76	2.67	0.102
PWP_2	-.27 (.17)	-.64 (.24)	.15 (.33)	-.61 (.43)	-.33 (.12)	.72	7.33	0.007
PWP_3	-.28 (.17)	-.54 (.24)	.18 (.33)	-.71 (.44)	-.31 (.12)	.73	6.34	0.012
AG	--	--	--	--	-.28 (.12)	.76	5.14	0.023
WLW	-.28 (.17)	-.55 (.23)	-.17 (.32)	-.45 (.40)	-.30 (.16)	.74	3.39	0.065
PC	-.28 (.28)	-.42 (1.28)	.35 (2.20)	-.25 (7.59)	-.28 (.28)	.76	1.03	0.309
PC_m	-.28 (.28)	-.50 (.49)	.14 (.91)	-.70 (2.55)	-.31 (.24)	.73	1.74	0.187

*$\hat{RR}$ = Estimated Relative Risk

Table 14
Consensus Trial
Analysis of Age Effect on Hospital Admissions
Adjusting for Treatment, History of MI and Diabetes Status

Model	Parameter Estimates[1] (standard error) Admission Number: 1	2	3	4	All	$\hat{RR}$[2]	Chi-Square	P-value
Cox	.09 (.10)	--	--	--	--	1.09	0.88	0.348
PWP_2	.08 (.10)	.19 (.15)	-.11 (.18)	.23 (.25)	.09 (.07)	1.09	1.58	0.208
PWP_3	.08 (.10)	.11 (.15)	-.27 (.17)	.46 (.31)	.06 (.07)	1.06	.59	0.441
AG	--	--	--	--	.08 (.07)	1.08	1.32	0.251
WLW	.09 (.09)	.19 (.17)	-.03 (.22)	.37 (.28)	.10 (.09)	1.11	1.28	0.257
PC	.09 (.13)	.11 (.67)	-.19 (1.03)	.50 (2.95)	.10 (.13)	1.11	0.52	0.471
PC_m	.09 (.13)	.12 (.30)	-.28 (.50)	.65 (1.14)	.08 (.12)	1.08	0.39	0.532

[1]Estimates represent change in log hazard per 10 year change in age

[2]$\hat{RR}$ = Relative Risk

Table 15
Consensus Trial
Effect of Previous MI on Hospital Admissions
Adjusting for Treatment, Age and Diabetes Status

Model	Parameter Estimates (standard error) Admission Number: 1	2	3	4	All	$\hat{RR}$*	Chi-Square	P-value
Cox	.47 (.17)	--	--	--	--	1.60	7.68	0.006
PWP_2	.48 (.17)	.23 (.24)	.25 (.32)	.06 (.44)	.35 (.12)	1.39	7.99	0.005
PWP_3	.48 (.17)	.08 (.24)	.53 (.33)	-.27 (.44)	.33 (.12)	1.39	7.11	0.008
AG	--	--	--	--	.40 (.12)	1.49	10.39	0.001
WLW	.47 (.17)	.50 (.24)	.66 (.32)	.68 (.41)	.48 (.17)	1.62	8.08	0.004
PC	.47 (.27)	.03 (1.19)	.33 (2.06)	-.29 (6.46)	.47 (.27)	1.60	3.05	0.081
PC_m	.47 (.27)	.13 (.53)	.50 (.88)	-.42 (2.73)	.42 (.24)	1.52	3.09	0.079

* $\hat{RR}$ = Estimated Relative Risk

Table 16
Consensus Trial
Effect of Diabetic Status on Hospital Admissions
Adjusting for Treatment, Age and Previous MI

Model	Parameter Estimates (standard error) Admission Number: 1	2	3	4	All	$\hat{RR}$*	Chi-Square	P-value
Cox	.21 (.19)	--	--	--	--	1.32	1.12	0.290
PWP_2	.20 (.19)	.66 (.25)	.09 (.33)	.22 (.44)	.31 (.13)	1.36	5.49	0.019
PWP_3	.21 (.19)	.71 (.24)	.30 (.33)	.75 (.47)	.40 (.13)	1.49	9.34	0.002
AG	--	--	--	--	.49 (.13)	1.63	13.88	<0.001
WLW	.21 (.19)	.81 (.23)	.87 (.32)	.97 (.43)	.33 (.19)	1.39	3.08	0.079
PC	.21 (.32)	.64 (1.05)	.33 (2.26)	.91 (5.93)	.23 (.31)	1.26	0.56	0.454
PC_m	.21 (.32)	.69 (.48)	.32 (.83)	.74 (2.40)	.35 (.27)	1.42	1.71	0.191

* $\hat{RR}$ = Estimated Relative Risk

B. Hospitalization in patients with Congestive Heart Failure - SOLVD

i. Study Description

The SOLVD treatment intervention trial studied the effect of the angiotensin-converting-enzyme (ACE) inhibitor enalapril on mortality and hospitalization in patients with chronic heart failure (SOLVD 1991). Over a 33-month period, a total of 2569 patients were enrolled in the study and randomly assigned to receive either enalapril (n = 1285) or placebo (n = 1284). The patients enrolled in this study were symptomatic, had left ventricular ejection fractions $\leq$ 0.35 and were currently receiving conventional treatment, other than ACE inhibitors, for CHF. The study was double-blind.

The average length of follow-up was 41 months. There were 950 patients in the placebo group (74%) and 893 patient in the enalapril group (69%) who were hospitalized at least once during the study (p = 0.006). Although SOLVD patients tended to have less severe heart failure than the patients from CONSENSUS, the hospital admission percentages were slightly higher in this group due to the longer duration and lower mortality in this trial.

Because the study was so large and the treatment effect was so strong, there was more than adequate power to detect the effect of treatment on time to first hospitalization. In an attempt to demonstrate whether there is any improvement in power by considering time to all events, the following analysis was performed on the subgroup of African-American patients. This is an interesting sub-population to study because African-Americans are known to respond differently than whites to some types of cardiovascular agents, including ACE inhibitors.

ii. Results - Treatment Effect

Table 17 summarizes the important baseline characteristics as well as the hospitalization history of the African-American patients included in this analysis. There were 187 patients who received placebo and 207 who received enalapril. The average age at baseline was 58 years for placebo patients and 56 years for the enalapril patients. Mean ejection fractions were similar for the two groups; placebo: 24.6% and enalapril: 24.3%. One hundred thirty-five patients in the placebo group (72%) and 139 patients in the enalapril group (67%) were hospitalized at least once during the entire duration of the trial. These percentages were slightly lower for this sub-population (2 percentage points in each treatment group) than for the entire study population, but the difference between the two treatment groups was similar.

As with the CONSENSUS hospitalization study, Cox's model for time to first hospitalization, AG, PWP_2, PWP_3, WLW, PC and PC_m were all used to analyze time to hospitalization, including up to four hospital admissions per subject (since the number of patients with more than four hospital admissions was small). In addition to the treatment effect, each method adjusted for the patient's age and ejection fraction at baseline. The results are summarized in Table 18. Parameter estimates and standard errors are shown for each event and combined overall events. Relative risk ($\hat{RR}$) was calculated as $\exp\{\hat{\beta}\}$.

The estimated treatment effect from Cox's analysis of time to first event, after adjusting for age and ejection fraction, was $\hat{\beta} = -0.23$ with a standard error of 0.12. Thus the estimated hazard for hospitalization for the enalapril group relative to the placebo group was 79%, or about

a 21% reduction in risk. With a chi-square of 3.66, the treatment difference was not quite significant ($p = 0.056$).

The PWP_2 model resulted in an estimated relative risk over all events of 77%, or a 23% reduction in risk for the enalapril group ($\hat{\beta} = -0.26$, $se = 0.08$). The first event parameter estimate from PWP_2 was $\hat{\beta} = -0.26$ ($se = 0.12$) which was the same as the overall estimate. This estimate differed from the Cox model because of the way the two models handle additional covariates. The estimated treatment effect for time to the second hospital admission was the same as the first event, but increased to a 30% reduction in risk at the third event($\hat{\beta} = -0.35$, $se = 0.20$, $\hat{RR} = 70\%$). By the fourth event, the estimated reduction in risk was only 10% ($\hat{\beta} = -0.11$, $se = 0.25$). As with the first example, standard errors of the parameter estimates increased with increasing event number.

The PWP_2 parameter estimates for this example can be interpreted as follows: Among patients who were hospitalized two or more times, there is a 29% reduction in hospitalization as measured from the start of the trial. But the interpretation of the individual event parameters is questionable for the reasons given in the previous example.

The chi-square from the overall test of a treatment effect is highly significant, $p = 0.002$. But as was previously pointed out, inference from the model is not valid. The covariance matrix for parameter estimates from the WLW model, showing non-zero off-diagonal elements is included in Appendix II, Table 2.

PWP_3 differs from PWP_2 in that it models gap times instead of total times. The estimated relative risk for the effect of treatment on time to hospital admissions combined overall events was 79% ($\hat{\beta}$ =-0.24, se=0.08). For the first event, the estimated relative risk was similar to the estimates for the Cox model and PWP_2, a relative risk of 77% ($\hat{\beta}$ =-0.24, se=0.12); *i.e.* a 23% reduction in risk. For the second event, the relative risk of hospitalization was 80% ($\hat{\beta}$ =-0.22, se=0.16), representing a 20% reduction in risk, slightly smaller than for the first event and smaller than the corresponding estimate for the PWP_2 model. By the third event, the estimated reduction in risk for hospitalization in the enalapril group had increased to 30%, $\hat{RR}$ =70% ($\hat{\beta}$ =-0.36, se=0.20) and by the fourth event, the relative risk was close to one, $\hat{RR}$ =96% ($\hat{\beta}$ =-0.04, se=0.25).

As pointed out before, parameter estimates from the individual events from PWP_3 are the most meaningful. For all methods that provide parameter estimates for individual events, the meaning of the parameter for time to first hospital admission is the same as the regular Cox model. For the subsequent events, the PWP_3 parameter represents the effect of treatment on prolonging the time between events. In this example, the PWP_3 parameter from the third event represents a 30% reduction in the risk of hospitalization, as measured from the second admission.

The chi-square test for the null hypothesis of no treatment effect overall events resulted in a p-value of 0.004. As in the previous example the validity of this test depends on independent gap times and may be exaggerating the evidence of a treatment effect by treating individual events as independent. Inference concerning the individual events is valid

and indicates a significant reduction in risk of the first hospitalization for the enalapril group.

The next method is AG, which provides parameter estimates and results overall events only. The estimated relative risk was 73%, $\hat{\beta} = -0.31$ with a standard error of 0.08. AG models hospitalization rate and the parameter represents the overall reduction in risk of hospitalization for all patients. The associated chi-square was highly significant, but inference from the AG model is questionable in the presence of correlated event times.

Based on the WLW model, the estimated treatment effect over all events was $\hat{RR} = 77\%$, ($\hat{\beta} = -0.26$, se = 0.12). The parameter estimate for the first event was the same as the Cox model and similar to the PWP_3, but for the next three events the estimated relative risks were much more negative, indicating greater risk reduction, than the corresponding PWP_3 estimates. The WLW estimates were $\hat{RR} = 69\%$ ($\hat{\beta} = -0.37$, se = 0.16) for event 2, $\hat{RR} = 56\%$ ($\hat{\beta} = -0.57$, se = 0.20) for event 3 and $\hat{RR} = 51\%$ ($\hat{\beta} = -0.67$, se = 0.26) for event 4. The increase in standard errors with event number was similar to the PWP models.

As mentioned before, since WLW models total times and includes all patients in the risk set for an event, regardless of the number of previous events, the meaning of the individual event parameters is questionable. When all patients are at risk for all events, as long as they remain in the study at time t, information from patients who were never hospitalized, which should only be included in the estimation of the first event effects, is included in the parameter estimates for each of the individual events.

This study clearly demonstrates the impact of WLW's risk set and use of total times on the interpretation of model parameters. Assuming that the PWP_3 parameters for the individual events are the most meaningful, it is obvious that the WLW parameter estimates are inflated for each event after the first. For example, the estimated 43% reduction in risk of a third hospital admission from the WLW model is much greater than the estimated 30% reduction from PWP_3. Again, the two reasons for this difference are the following 1) The risk set definition makes a big difference in this example. The crude reduction in third hospitalization rate was 34% for all patients, WLW risk set, but only 17% among patients with at least 2 hospitalizations, the PWP_3 risk set. 2) By modelling total time, the effects of enalapril on the first 2 events tend to carry over for the WLW estimate of the third and fourth event effects, and in this case creating a cumulative treatment effect by the latter events.

In spite of problems with the individual event estimates, it is interesting that the estimate for a treatment effect over all events from this model appeared to be reasonable relative to the PWP_3 estimate.

Ignoring the parameter estimates from individual events, WLW has the best statistical properties of all methods based on the simulation results in the previous chapter. WLW does adjust for correlations between event times and appears to maintain significance levels close the nominal level, even under heavy censoring. The associated Wald chi-square for an overall treatment effect based on the WLW model was 4.68, $p = 0.030$. Recall that the p-value from the Cox first event analysis was 0.056 and therefore using the WLW model, after adjusting for the correlation among event times, there is still enough additional information from the second

through fourth hospital admissions to declare the treatment effect statistically significant.

Results for PC and PC_m are shown in the table, but not discussed in detail. Individual event and averaged parameter estimates for PC_m and PC were similar to those observed for PWP_3. As shown in the Table 18 these methods had problems with increasing standard errors at the later events. This is probably due to a combination of the small sample size and high proportion of censored patients.

iii. Results - Additional Covariates

In the SOLVD example the effects of age and ejection fraction at baseline were also examined. Table 19 summarizes the analysis of the effect of age on hospital admissions controlling for treatment and baseline ejection fraction. The parameter estimates represent the change in log hazard or increase in risk per 10 year change in age. Based on WLW, there appeared to be a 12% increase in the risk of hospitalization per 10 year increase in age, $p=0.050$. This was the same as the estimated risk from the Cox model. The PWP_3 individual event estimates confirmed that the effect of age was present at the first event only.

Results from the analyses of the effect of baseline ejection fraction on risk of hospital admissions in models that included treatment and age are summarized in Table 20. The parameter estimates represent the change in risk per 10 percentage point change in ejection fracton. Although we might expect lower injection fractions to be associated with an increased risk of hospitaliztion, there was no evidence of an association in this study.

iv. Summary

Conclusions from this analysis with respect to the effect of enalapril on risk of hospitalization in patients with heart failure should be based on inference from the WLW method. These results indicate that enalapril significantly reduces the risk of hospitalization in patients with heart failure. The estimated relative risk is approximated 77%, representing a 23% risk reduction. This conclusion would be not possible based on the Cox model, since the difference between treatments in time to first event was not significant.

Age appears to be associated with an increased risk of hospitalization.

Although mortality was not as high in this study as in the CONSENSUS study, the problem of death as a competing risk for the hospitalization endpoint still exists to some extent. Again if the sicker patients are dying before they have a chance to be hospitalized, and the treatment has a positive effect, then the bias from this type of censoring will most likely be conservative.

Table 17
SOLVD - Treatment Trial
Subject characteristics and Hospitalization History for African-American Patients

Variabl	Placebo (n=187)	Enalapril (n=207)
Age (yr)		
mean	57.9	55.9
s.d.	11.0	11.4
Ejection Fraction (%)		
mean	24.6	24.3
s.d.	7.1	7.0
Patients Hospitalized		
Number (%) with:		
1 admission	135 (72.2%)	139 (67.1%)
2 admissions	89 (47.6%)	78 (37.7%)
3 admissions	62 (33.2%)	45 (21.7%)
4 admissions	40 (21.4%)	27 (13.0%)
5 admissions	25 (13.4%)	16 (7.7%)
6 admissions	15 (8.0%)	12 (5.8%)

Table 18
SOLVD - Treatment Trial
Analysis of Treatment effects on Hospital Admissions
for African American Patients
Adjusting for Age and Ejection Fraction at Baseline

Model	Parameter Estimates (standard error) for Admission Number: 1	2	3	4	All	$\hat{RR}$*	Chi-Square	P-value
Cox	-.23 (.12)	--	--	--	--	.79	3.66	0.056
PWP_2	-.26 (.12)	-.26 (.16)	-.34 (.20)	-.11 (.25)	-.26 (.08)	.77	10.01	0.002
PWP_3	-.24 (.12)	-.22 (.16)	-.36 (.20)	-.04 (.25)	-.24 (.08)	.79	8.43	0.004
AG	--	--	--	--	-.31 (.08)	.73	15.00	<0.001
WLW	-.23 (.12)	-.37 (.16)	-.57 (.20)	-.67 (.26)	-.26 (.12)	.77	4.68	0.030
PC	-.23 (.19)	-.19 (.70)	-.40 (1.34)	-.04 (2.72)	-.23 (.18)	.79	1.64	0.201
PC_m	-.23 (.19)	-.25 (.28)	-.36 (.37)	-.02 (.68)	-.27 (.15)	.76	3.22	0.073

*$\hat{RR}$ = Relative Risk overall events.

Table 19
SOLVD - Treatment Trial
Analysis of Effect of Age at Baseline on Hospital Admissions for African American Patients
Adjusting for Treatment and Baseline Ejection Fraction

Model	Parameter Estimates[1] (standard error) for Admission Number: 1	2	3	4	All	$\hat{RR}$[2]	Chi-Square	P-value
Cox	.11 (.05)	--	--	--	--	1.12	4.84	0.029
PWP_2	.11 (.05)	-.02 (.07)	-.10 (.09)	-.02 (.11)	.02 (.04)	1.02	0.34	0.559
PWP_3	.11 (.05)	.00 (.07)	.07 (.09)	.02 (.11)	.06 (.04)	1.06	2.71	0.100
AG	--	--	--	--	.08 (.04)	1.08	4.32	0.038
WLW	.11 (.05)	.10 (.07)	.04 (.09)	.03 (.11)	.11 (.06)	1.12	3.84	0.050
PC	.11 (.08)	.04 (.30)	.14 (.71)	.00 (1.19)	.11 (.08)	1.12	1.60	0.206
PC_m	.11 (.08)	.01 (.13)	.04 (.19)	.01 (.31)	.08 (.07)	1.08	1.35	0.246

[1]Estimates represent change in log hazard per 10 year change in age
[2]$\hat{RR}$ = Relative Risk

Table 20
SOLVD - Treatment Trial
Analysis of Effect of Baseline Ejection Fraction on Hospital Admissions
for African American Patients
Adjusting for Treatment and Age at Baseline

Model	Parameter Estimates[1] (standard error) for Admission Number: 1	2	3	4	All	$\hat{RR}$[2]	Chi-Square	P-value
Cox	-.06 (.08)	--	--	--	--	.94	0.47	0.493
PWP_2	-.05 (.08)	-.05 (.11)	.17 (.14)	.17 (.18)	.01 (.06)	1.01	0.05	0.823
PWP_3	-.05 (.08)	-.12 (.11)	.09 (.13)	-.09 (.18)	-.05 (.06)	.95	0.68	0.411
AG	--	--	--	--	-.05 (.06)	.95	0.62	0.430
WLW	-.06 (.08)	-.09 (.11)	.00 (.14)	.10 (.18)	-.06 (.08)	.94	0.53	0.466
PC	-.06 (.13)	-.10 (.54)	.07 (1.01)	-.11 (1.95)	-.06 (.12)	.94	0.22	0.640
PC_m	-.06 (.13)	-.10 (.20)	.10 (.28)	-.04 (.50)	-.05 (.10)	.95	0.23	0.633

[1]Estimates represent change in log hazard per 10 percentage point change in ejection fraction.

[2]$\hat{RR}$ = Relative Risk

C. Myocardial Infarction - SOLVD Treatment and Prevention Trials

i. Study Description

The third example combines the SOLVD treatment and preventions trials to assess the effect of the ACE inhibitor enalapril on myocardial infarction (MI) and unstable angina in patients with low ejection fractions (SOLVD 1992). The SOLVD treatment trial was described in the first example and the prevention trial was similar in that it studied patients with ejection fractions $\leq$ 0.35%; however, the prevention trial patients were asymptomatic and were not currently receiving drug therapy for CHF (SOLVD 1992). Both studies were double-blind, placebo controlled.

Raised blood pressure and increased renin levels have both been associated with an increased incidence of MI. ACE inhibitors may reduce the risk of MI through multiple beneficial effects such as lowering blood pressure and reducing angiotensin-II levels.

ii. Results

Combining the two trials, a total of 3395 placebo patients and 3387 enalapril patients were included in the analysis. Table 21 summarizes the mean age, ejection fraction and the MI history of patients included in the analysis. The mean age in both treatment groups was approximately 59 years and the mean ejection fraction was 27%. Three hundred seventy one patients in the placebo group (10.9%) and 294 patients in the enalapril group (8.7%) suffered at least one MI.

The multiple event procedures were used to analyze time to MI, including up to two MIs per subject. The number of subjects with more than 2 MIs was extremely small. Because of the large number of patients, inclusion of additional covariates, other than treatment, was computationally impractical. Attempts to include additional covariates resulted in insufficient memory problems.

Results from the analyses are summarized in Table 22. Based on the Cox model, the risk of a first MI was reduced by 24% in the enalapril group compared to the placebo group, ($\hat{\beta}$ =-0.27, se=0.08, p<0.001).

The PWP_2 model resulted in an estimated reduction in risk of 19%, ($\hat{\beta}$ =-.21, se=0.07), combining both events. The first event estimate and standard error were identical to the Cox model, which is expected since there were no additional covariates in the model. There appeared to be a small increase in the risk of a second MI for the enalapril group, $\hat{RR}$ =116%, or an increased risk of 16%. The standard errors for the estimates increased from the first to the second event.

The PWP_2 second event parameter represents the increased risk in MI among patients who have already had an MI, calculated from the start of the study.

The chi-square from the overall test of a treatment effect was significant, p=0.003; however inference should not be made from this model.

The next method applied was PWP_3, which resulted in the same overall and first event parameter estimates as PWP_2. The second event

parameter from PWP_3 represented an increased risk of 9% in the enalapril group; however, the size of the risk was negligible relative to the standard error ($\hat{\beta} = 0.09$, se = 0.18).

PWP_3 provides the most meaningful parameter estimates. In this example, the second event estimate indicated that while enalapril reduced the risk of a first MI, there was no improvement in risk of a second MI among the patients who have had an MI.

The chi-square test for the null hypothesis of no treatment effect over the two events resulted in a p-value of 0.003. This test is not valid unless there is no correlation between the two gap times. Results for hypotheses about the individual event parameters from this model are valid. Inference about the first event is the same as the Cox model and will always be the same as long as there are no additional covariates (or the additional covariates are estimated separately over strata). So from the Cox model, we found a reduction in risk of MI in the enalapril group. PWP_3 showed that there was no reduction in risk of a second event for patients who have had an MI.

The estimated relative risk from the AG model was 77%, representing a 23% reduction in the risk of MI for the enalapril group. ($\hat{\beta} = -0.26$, se = 0.07). The associated chi-square was significant, $p < 0.001$; however, inference from the model requires independence among the events.

The estimated overall reduction in risk of MI obtained from WLW was 24% ($\hat{\beta} = -0.27$, se = 0.08). This is the same as the Cox first event parameter estimate and standard error. The first event parameter from

WLW was also the same as the parameter estimates from both PWP_2 and PWP_3. The estimate for the second event represented a relative risk of the same size as the first ($\hat{RR} = 76\%$); whereas the estimate from PWP_3 suggested the treatment effect was in the opposite direction ($\hat{RR} = 109\%$). Standard errors for the two estimates were the same. This is another example of how the risk set and definition of event times used by the WLW method can result in misleading parameter estimates.

In addition to the parameter estimates, the associated Wald chi-square for an overall treatment effect based on the WLW model was identical to the test from Cox's model for time to first MI. Thus there appears to be very little additional information with respect to the effect of treatment on MI after the first event.

Results for PC and PC_m are shown in the table. Individual event and averaged parameter estimates for PC_m and PC were similar to those observed for PWP_3 .The standard errors were much larger at the second event.

iv. Summary

Enalapril reduces the risk of MI. This conclusion is based on results from WLW, which considers two MI events but are identical to the results for the first event analysis. Furthermore, examining the treatment effect for the second event from PWP_3, suggests that all the benefit comes from the first event information. Enalapril reduces the risk of a first MI. Among patients who have had an MI, there is no evidence that enalapril reduces the risk of a second. Death as a competing risk should also be acknowledged in this example.

Table 21
SOLVD - Treatment and Prevention Trials Combined
Subject Characteristics and Myocardial Infarction History

Variable	Placebo (n = 3395)	Enalapril (n = 3387)
Age (yr)		
mean	59.4	59.3
s.d.	10.2	10.2
Ejection Fraction (%)		
mean	27.0	27.0
s.d.	6.3	6.3
Myocardial Infarctions		
Number (%) with:		
1 MI	371 (10.9%)	294 (8.7%)
2 MIs	74 (2.2%)	59 (1.7%)
3 MIs	8 (<0.1%)	8 (<0.1%)
4 MIs	2 (<0.1%)	2 (<0.1%)

Table 22
SOLVD- Treatment and Prevention Trials Combined
Analysis of Treatment effect on Myocardial Infarction

Model	Parameter Estimates (standard error) for MI Number: 1	2	All	$\hat{RR}$*	Chi-Square	P-value
Cox	-.2 (.08)	--	--	.77	12.07	<0.001
PWP_2	-.27 (.08)	.11 (.18)	-.21 (.07)	.81	8.58	0.003
PWP_3	-.27 (.08)	.09 (.18)	-.21 (.07)	.81	8.88	0.003
AG	--	--	-.26 (.07)	.77	13.51	<0.001
WLW	-.27 (.08)	-.27 (.18)	-.27 (.08)	.76	12.07	<0.001
PC	-.27 (.35)	.11 (2.41)	-.28 (.35)	.76	0.64	0.423
PC_m	-.27 (.35)	.09 (1.33)	-.31 (.33)	.70	0.84	0.359

*$\hat{RR}$ = Relative Risk overall events.

Chapter V
DISCUSSION AND DIRECTIONS FOR FUTURE RESEARCH

A. Summary

This research focused on the evaluation of six methods for analyzing time-to-event data in clinical trials where there are multiple failures of the same type.

Chapter I introduced the problem of recurrent events in survival analysis and outlined several of the important features that distinguish the different methods, including assumptions about dependency within subjects, the definition of event time and the different risk sets used. An overview of the research questions of interest was also presented.

Chapter II gave a broad review of the survival analysis approach to time-to-event data, emphasizing Cox's semi-parametric proportional hazard's model. The chapter also contained an overview of several approaches to analyzing multivariate data in general. Methods specifically designed for analyzing survival studies with multiple outcomes were reviewed in depth. Six models were chosen for evaluation and specific features of the models that might be expected to affect their statistical properties were discussed. All models were related to Cox's model. Three methods, AG, PWP_2, and PWP_3, made explicit assumptions about the dependency structure among event times within individuals, while the remaining methods, WLW, PC and PC_m, adjusted for the correlations in the covariance estimates. The impact of the methods' different risk sets and

definitions of event times on interpretation of model parameters was also discussed. In the types of examples analyzed as part of this research such as hospitalization admissions and occurrence of myocardial infarction, restricting the risk set to patients who have had the previous event and modelling time since previous event rather than time since study start were more intuitive than the other approaches. In these examples the meaning of the individual event parameters from the WLW marginal approach that modelled total time using an unrestricted risk set were questionable, while the parameters from PWP_3 and PC_m, which restrict the risk set and model gap times were more meaningful. It was acknowledged that there may be examples such as the occurrence of tumors or post-surgical infections where the alternate definitions of risk set and event time are reasonable but the fact that both definitions will introduce correlations among event times was noted.

Chapter III began with an outline of the strategy for the simulation analysis and discussed the questions the simulations were designed to answer. To evaluate true significance levels of the six methods identified in Chapter II, the simulations were run varying the following parameters: 1) degree of dependence between event times, 2) number of failure events, 3) sample size, 4) number of covariates and 5) degree of censoring. The methods were also compared with respect to power of detecting different magnitudes of treatment effects, as a function of the dependence between event times and the amount of censoring.

The simulation results showed that the true significance level of PWP_2 is always inflated. For the PWP_3 and AG methods, the asymptotic properties of the tests were excellent unless there was correlation among events. Under these conditions the true significance levels were inflated.

The WLW method did not have as good asymptotic results as the PWP_3 and AG methods when independent event times were generated, but performed better than any of the other methods when event times were correlated. The small sample properties of PC_m were slightly worse than WLW's. The significance levels of PC suggested small sample problems.

WLW was the only method that maintained significance levels when censoring was introduced. The significance level of all other methods dropped dramatically as the amount of censoring increased.

AG and PWP_3 were always more powerful than WLW and PC_m, but the tests were only valid under conditions of independent event times. AG and PWP_3 would be appropriate methods for applications where the event times were known to be independent, but this is often an unlikely situation. PC_m was more powerful than WLW except in the situations where data were censored. Thus, PC_m would be an appropriate alternative for situations where events were not censored, or the degree of censoring was small. Again this is atypical of survival-type studies.

The power of the WLW test was compared to a Cox analysis in simulations of four events. WLW was less powerful than Cox when the treatment effect was present at the first event only, but became more powerful when at least half the effect from the first event was present at the second event. The improvement in relative power increased as stronger effects persisted over more events.

WLW had the best statistical properties in situations where event times might were correlated and censoring was present. The method did have drawbacks: It was not very powerful relative to the other methods and interpretation of covariate effects for individual events was questionable.

Interpretation and inference about the parameters from individual events estimates from the PWP_3 and PC_m models are the most meaningful.

Chapter IV presented results from the analysis of three examples. The effect of enalapril relative to placebo was examined in patients with congestive heart failure with respect to the risk of hospital admissions in the CONSENSUS study and in the sub-group of African-American patients in the SOLVD Treatment study. All six methods were used to analyze both studies but WLW was considered the best method for making inferences regarding a treatment effect combining information from all events.

In the CONSENSUS study the p-value for a significant treatment difference was lower with the WLW method that considered up to four admissions than with the Cox first event analysis, but was still not quite significant. Overall parameter estimates were consistent among the six methods; however, there was wide variation in the standard errors and the resulting p-values. Individual event parameter estimates from WLW were questionable in that they showed carryover of first event information to subsequent event estimates.

The second example was the sub-group analysis of African-American patients from the SOLVD treatment study. The WLW analysis of the first four hospital admissions showed that enalapril significantly reduced the risk of hospitalization, whereas the difference between treatments in the time-to-first event was not significant. Again, the overall parameter estimates were similar for all methods.

The third example was a study of the occurrence of myocardial infarction in the combined SOLVD treatment and prevention trials. In this example

the results of the WLW analysis that considered two events was identical to the Cox analysis of first MI.

The problem of death as a competing risk for hospital admission or occurrence of MI was discussed for each example.

The examples were important in demonstrating how a multiple event analysis might be beneficial with respect to a first event analysis. They illustrated how consistent the parameter estimates were, particularly for the treatment effect overall events. Finally, the examples showed how the risk set and event time definitions used by WLW may create misleading model parameters at later events.

B. Directions for Future Research

Among the six methods evaluated, WLW appeared to be the best for analyzing multiple time-to-event data but WLW also has two major drawbacks: The method has low power and parameter estimates for treatment effects for individual events are not meaningful.

Lin (1993) proposed developing robust variance estimators, similar to the WLW variance, for other methods such as PWP_3. Once developed, methods based on these new covariance matrices could be investigated to determine whether desirable statistical properties with respect to the size of the tests show improvement over the WLW method. If the PWP_3 type tests using a robust variance estimator was shown through simulations to be valid, the power of these methods could be investigated. A valid method with improved power and meaningful individual event parameters would have advantages over WLW. But since the proposed covariance estimator is a variation of the WLW estimator,

a gain in power may not be seen with these new methods.

The small sample problems seen with PC and PC_m as evaluated here may have been due to how these methods were implemented. The original method as published by Pepe and Cai (1993) was designed to analyze first events and all subsequent recurrent events as a group. Details for using the method for an analysis that considered each individual event were not given. A different approach to weighting the estimates and reducing the standard errors at later events should be explored.

Another possible area for future research would be to develop a randomization test based on the PWP_3 or PC_m parameter estimates. The advantages of randomization tests are that no assumptions need to be made regarding survival distribution or the correlation structure of recurrent failures within patients. If the total number of patients in the study is N, where $N = n_1 + n_2$, then the number of possible ways of allocating N patients in the study with n_1 in treatment group 1 is $\left(\begin{bmatrix} N \\ n_1 \end{bmatrix}\right)$. For each allocation one computes the summary statistic of interest, then the one tailed p-value is the proportions of permutations with a statistic as extreme as or more extreme than the observed statistic. In practice the number of permutations $\left(\begin{bmatrix} N \\ n_1 \end{bmatrix}\right)$ is too large and Monte Carlo sampling is used to randomly select a fixed number from the permutations (Freedman and Byar 1989).

Confidence intervals can be obtained from the sampling distribution.

In order to define a randomization test, it is first necessary to select an

appropriate test statistic upon which such a test should be based. One approach would be to use a single estimate that is a weighted average of the individual event estimates from the PWP_3 or PC_m methods. The weights should account for fewer patients at risk and fewer expected events at the later events.

An additional area for further research would be investigating ways of incorporating informative censoring into the analyses. Methods incorporating this information would be especially useful in applications where competing risk can produce substantial bias or where there is allot of patient withdrawal due to lack of therapeutic effects or toxicity of a therapeutic agent. Development of such methods could be based on Pepe's (1991) approach to analyzing events with dependent risks, but allowing for recurrent endpoints of a single type. Intention-to-treat and dropout analysis principles could also be incorporated.

In summary, through simulations and applied examples this research clearly demonstrates the need for statistical methods which address multiple event problems in a valid and intuitive manner. Although WLW is shown to be the most valid method for making inferences across all events in the presence of correlated event times and censoring, the low power and lack of interpretation of individual event results reduce its applicability. Other methods provide better interpretation of the analyses of individual events but have problems with valid inference across all events when event times are correlated. Several other methods exhibit small sample instability. Further research in these areas would greatly aid the applied statistician seeking methods which work well under a variety of conditions that are typical of clinical trials. However, the current research provided a benchmark of performance and a strategy for evaluation of these methods and should be a useful tool for statisticians

who are investigating appropriate methods for time-to-event analysis where multiple events can occur.

BIBLIOGRAPHY

Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes. *Annals of Statistics* **6**: 701-726.

Andersen, P.K., Borgan, O., Gill, R. and Keiding, N. (1982). Linear Nonparametric Tests for Comparison of Counting Processes, with Applications to Censored Survival Data. *International Statistical Review* **50**: 219-258.

Andersen, P.K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* **10**: 1100-1120.

Abu-Libdeh, H., Turnbull, B. W, and Clark, L. C. (1990). Analysis of Multi-Type Recurrent Events in Longitudinal Studies, Application to a Skin Cancer Prevention Trial. *Biometrics* **46**: 1017-1034.

Breslow, N. E. (1972). Contribution to Discussion of Paper by D. R. Cox. *J. R. Statist. Soc. B* **34**: 187-220.

Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics* **30**: 89-99.

Breslow N. R. (1975). Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review* **43**: 45.

Cheuvart, B. (1988). A Nonparametric Model for Multiple Recurrences. *Appl. Statist.*, **37**(2): 157-168.

Clayton, D. and Cuzick, J. (1985). Multivariate Generalizations of the Proportional Hazards Model. *J. R. Statisti. Soc. A* **148**(2): 82-117.

The CONSENSUS Trial Study Group (1987). Effect of Enalapril on Mortality in Severe Congestive Heart Failure: Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS). *N. Engl. J. Med* **316**: 1429-35.

Cox, D. R. (1972). Regression Models and Life-Tables (with discussion). *J. R. Statist. Soc. B* **34**: 187-220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika* **62**: 269-276.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival Models and Data Analysis*. Wiley: New York.

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

Freedman, L., Sylvester, R. and Byar, D. P. (1989). Confidence Limits to Analyze Repeated Events Data From Clinical Trials. *Controlled Clinical Trials* **10**: 129-141.

Gail, M. H., Santner, T.J. and Brown, C. C. (1980). An Analysis of Comparative Carcinogenesis Experiments Based on Multiple Times to Tumor. *Biometrics* **36**, 255-266.

Harrell, F. E. (1992). *Survival and Risk Analysis*. Copyrighted but unpublished class notes.

Hougaard, P. (1986). A Class of Multivariate Failure Time Distributions. *Biometrika* **73**: 671-678.

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

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's Regression and Life Model. *Biometrika* **60**: 267-278.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**: 457-481.

Lawless, J. F. (1987). Regression Methods for Poisson Process Data. *Journal of the American Statistical Association* **82**(399): 808-815.

Lee, E. T. (1980). *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications: Belmont California.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations. *Survival Analysis: State of the Art*, Ed. J. P. Klein and P. K. Goel: 237-247. Kluwer Academic Publishers, Dordrecht. Publications: Belmont California.

Liang, K.Y. and Zeger, S. L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika* **73**: 13-22.

Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association* **84**(408): 1074-1078.

Lin, D. Y. (1993). Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach. *Statisti. in Med*: Submitted.

Lin, D. Y. (1993). Mulcox2: a General Computer Program for the Cox Regression Analysis Multivariate Failure Time Data. *Computer Methods and Programs in Biomedicine*: Submitted.

Lin, J. S. and Wei, L. J. (1992). Linear Regression Analysis for Multivariate Failure Time Observations. *Journal of the American Statistical Association* **87**(420): 1091-1097.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall: New York.

Miller, R.G., Jr. (1981). *Survival Analysis*. Wiley: New York.

Oakes, D. (1989). Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association* **89**(406): 487-493.

Pepe, M. S. (1991). Inference for Events With Dependent Risks in Multiple Endpoint Studies, *Journal of the American Statistical Association* **86**(415): 770-778.

Pepe, M. S. and Cai, J.(1993). Some Graphical Displays and Marginal Regression Analysis for Recurrent Failure Times and Time Dependent Covariates. *Journal of the American Statistical Association* **88**(423): 811-820.

Pepe, M. S. and Fleming T. R. (1991). Weighted Kaplan-Meier Statistics: Large Sample Theory and Optimality Considerations. *Journal of Royal Statistical Society*, Series B. **53**: 341-352.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981), On the Regression Analysis of Multivariate Failure Time Data. *Biometrika* **68**: 373-379.

Prentice, R. L. and Cai, J. (1992). Covariance and Survival Function Estimation Using Censored Multivariate Failure Time Data. *Biometrika* **79**: 495-512.

Prentice, R. L. and Zhao, L. P. (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Response. *Biometrics* **47**(3): 825-840.

The SOLVD Investigators (1991). Effect of Enalapril on Survival in Patients with Reduced Left Ventricular Ejection Fractions and Congestive Heart Failure. *N. Engl. J. Med* **325**: 293-301.

The SOLVD Investigators (1992). Effect of Enalapril on Mortality and the Developement of Heart Failure in Patients with in Asymptomatic Patients with Reduced Left Ventricular Ejection Fractions. *N. Engl. J. Med* **327**: 685-691.

Thall, P. F. (1988). Mixed Poisson Likelihood Regression Models for Longitudinal Interval Count Data. *Biometrics* **44**: 197-209.

Thall, P. F. and Lachin, J. M. (1988). Analysis of Recurrent Events: Nonparametric Methods for Random-Interval Count Data. *Journal of the American Statistical Association* **83**(402): 339-347.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association* **84**(408): 1065-1073.

Wei, L. H., and Johnson, W. E. (1985). Combining Dependent Tests with Incomplete Repeated Measurements. *Biometrika* **72**(2): 359-64.

Zeger S. L. and Liang K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* **42**: 121-130.

Appendix I

Table 1

Actual Correlations Between Event Times Associated with Generated Data For Gap Times and Total Times

Gap Times

Value of Depend	Correlations for All events
0	.00
0.5	.07
1	.22

Total Times

	Correlations between events:		
Depend	1 and 2	1 and 3	2 and 3
0	0.69	0.60	0.81
0.5	0.71	0.60	0.83
1	0.77	0.68	0.88

Appendix II

Covariance Matrix for Parameter Estimates for Treatment Effect

Table 1: CONSENSUS Trial Analysis of Hospital Admissions

Covariance Matrix From WLW

0.02769	0.02399	0.02306	0.02137
0.02399	0.05429	0.04900	0.04379
0.02306	0.04900	0.10195	0.08931
0.02137	0.04379	0.08931	0.16196

Covariance Matrix From PC_m

0.07895	0.02097	0.05250	0.18054
0.02097	0.24340	0.22765	0.71134
0.05250	0.22765	0.83680	1.29421
0.18054	0.71134	1.29421	6.52325

Table 2: SOLVD Trial Analysis of Hospital Admissions For African-Americans

Covariance Based on WLW

0.01464	0.01238	0.01206	0.01227
0.01238	0.02477	0.02404	0.02358
0.01206	0.02404	0.03962	0.03905
0.01227	0.02358	0.03905	0.06588

Covariance Matrix from PC_m

0.03475	-0.00103	0.00557	0.01546
-0.00103	0.07583	0.03519	0.06486
0.00557	0.03519	0.13693	0.15816
0.01546	0.06486	0.15816	0.46537