Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 8

11-1-2004

On A Simple Method For Analyzing Multivariate Survival Data Using Sample Survey Methods

Pingfu Fu
Case Western Reserve University, pxf16@case.edu

J. Sunil Rao
Case Western Reserve University, sunil@hal.cwru.edu

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 345-356

1538 – 9472/04$95.00

On A Simple Method For Analyzing Multivariate Survival Data Using Sample Survey Methods

Pingfu Fu J. Sunil Rao
Department of Epidemiology and Biostatistics
Case Western Reserve University

A simple technique is illustrated for analyzing multivariate survival data. The data situation arises when an individual records multiple survival events, or when individuals recording single survival events are grouped into clusters. Past work has focused on developing new methods to handle such data. Here, we use a connection between Poisson regression and survival modeling and a cluster sampling approach to adjust the variance estimates. The approach requires parametric assumption for the marginal hazard function, but avoids specification of a joint multivariate survival distribution. A simulation study demonstrates the proposed approach is a competing method of recent developed marginal approaches in the literature.

Key words: sampling; design effect; survival analysis; clustered data

Introduction

Clustered survival events can occur in a number of ways. The form receiving considerable attention has been the scenario of when an individual is subject to experiencing repeat events (recurrent or multiple-type) over time. An illustration of this is the case where a child is diagnosed with chronic lung disease (CLD) for a period of time. The disease may or may not resolve. If resolution occurs, the child is susceptible to repeat occurrences of CLD over time (Norton, et. al., 2001). The time to the start of each CLD episode can be thought of a series of clustered events where the clustering unit is the child.

Pingfu Fu is an Assistant Professor of Biostatistics. E-mail: pxf16@case.edu. J. Sunil Rao is Associate Professor of Biostatistics, Department of Epidemiology & Biostatistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106. E-mail: sunil@hal.cwru.edu.

There have been a number of different methods proposed to handle inference in this situation. These include Andersen and Gill (AG) model (1982), Prentice, Williams and Peterson (PWP) model (1981), and Wei, Lin and Weisfeld (WLW) model (1989). In AG model, each subject is treated as a multi-event counting process with essentially independent increments; PWP model is a conditional approach; and WLW model is marginal method, in which one obtains the estimated coefficients, ignoring correlation, followed by fix of the variance of estimated coefficients.

More recently, Segal and Neuhaus (1993) showed how to use Poisson regression techniques to analyze such data. Their method made use of generalized estimating equation (GEE) machinery (Liang & Zeger, 1986) for doing point estimation. Robust inference was handled by using sandwich estimators for variance estimates of estimated regression parameters. In all of these applications, much of which has recently become widely available (Therneau & Grambsch, 2000) and can be fitted by major statistical software, such as SAS (SAS Institute, Cary, NC) and Splus (Insightful Corp., Seattle, Washington).

Survey sampling is another area where clustered events are quite common. The design effects approach, which is based on sample

survey techniques, has been used for analyzing such data. Design effect represents the estimated inflation in the variance of estimated coefficients due to correlated observations in each cluster (Rao and Scott, 1999). In order to account for the correlation among observations within each cluster, we can either transform the data by a design effect and apply standard methods afterwards assuming independence, or apply standard methods assuming independence, and then adjust the variances of the estimates by design effects. Work in non-survival setting includes that by Rao and Scott (1992, 1999) and Bieler and Williams (1995). In this paper, we are going to use the design effect approach under the survival analysis-Poisson regression and show how the design effects method can very simply handle clustered survival events, too.

Our method is similar to Segal and Neuhaus's approach in terms of the variance estimate - both use a sandwich estimator, but differ with respect with in the "filling". It's well-known that the Liang-Zeger's GEE application of quasi-likelihood on which Segal and Neuhaus's is based is essentially a special case of Binder's method (Binder, 1983) applied to with-replacement cluster sampling. Paik (1988) has shown that the GEE methods can lead to considerably biased parameter estimates in small sample settings. This is part of the motivation for the alternative approach we propose. Our method is parametric, and marginal, thus, it sacrifices the semi-parametric specification of AG, PWP and WLW. However, it provides another platform using only regular Poisson regression to analyze multivariate survival data.

Multivariate survival data and GEE

Assume that we have a sample of failure time data represented by

$$\left\{ \begin{array}{l} \left(T_{ijk}, \delta_{ijk}, x_{ijk} \right) : i \\ \qquad\qquad = 1, \ldots, G, j \\ \qquad\qquad = 1, \ldots, m_i, k = 1, \ldots, n_{ij} \end{array} \right\}$$

where for observation k of individual j of treatment group i, T_{ijk} denotes a failure time, δ_{ijk} is an event indicator taking the value 1 if T_{ijk} is uncensored and 0 otherwise, and x_{ijk} is a p-dimensional vector of covariates. There are assumed to be m_i individuals within treatment group i and G treatment groups in total. Let $S(t)$, $f(t)$ and $\lambda(t)$ be the survival distribution, density function and hazard function respectively for random variable T where $t \geq 0$ is a generic survival time.

Following Segal and Neuhaus (1993), we assume that the marginal hazard function for the kth observation of the jth individual in the ith treatment group involves covariates x_{ijk} through Cox's proportional hazards model

$$\lambda_{ijk}(t) = \lambda_0(t) \exp(\beta' x_{ijk}),$$

where β is a p-dimensional vector of regression parameters, and $\lambda_0(t)$ is the baseline hazard function. Thus

$$f_{ijk}(t) = \lambda_0(t) \exp\{\beta' x_{ijk} - \Lambda_0(t) e^{\beta' x_{ijk}}\}$$

and

$$S_{ijk}(t) = \exp\{-\Lambda_0(t) e^{\beta' x_{ijk}}\},$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function. As in Segal and Neuhaus, we depart from the standard Cox proportional hazards model which does not assign a parametric form for $\lambda_0(t)$.

Under the standard assumption of independent censoring, the likelihood for the kth observation of jth individual in the ith treatment group is

$$\begin{aligned} L_{ijk}(\alpha, \beta) &= f_{ijk}(t)^{\delta_{ijk}} \left(S_{ijk}(t)\right)^{1-\delta_{ijk}} \\ &= [\lambda_0(t) \exp(\beta' x_{ijk})]^{\delta_{ijk}} \exp(-\Lambda_0(t) e^{\beta' x_{ijk}}) \\ &= (\mu_{ijk}^{\delta_{ijk}} e^{-\mu_{ijk}})(\lambda_0(t) / \Lambda_0(t))^{\delta_{ijk}} \end{aligned} \quad (1)$$

where $\mu_{ijk}(t) = \Lambda_0(t) \exp(\beta' x_{ijk})$ and α are the parameters specifying the baseline survival distribution.

Because the δ_{ijk} takes on values of only 0 or 1, the first term in (1) can be thought of as a Poisson random variable with mean μ_{ijk}. A log-linear model for the hazard function implies a log-linear model for μ_{ijk} through

$$\log(\mu_{ijk}) = \log(\Lambda_0(t)) + \beta' x_{ijk}.$$

As mentioned earlier, we will give parametric form to $\lambda_0(t)$ or $\Lambda_0(t)$, say for example, by letting $\Lambda_0(t) = t$. Then *f(t)* is simply an exponential density with mean $\exp(-\beta' x)$, and maximum likelihood estimates for the regression parameters β can be found by fitting a Poisson regression model where response is the censoring variable with an offset $\log t_{ijk}$.

By assuming the independence of responses within each cluster, Segal and Neuhaus (1993) handle the clustering by fitting a corresponding GEE model (Liang and Zeger, 1986) and use robust sandwich estimators for inference on the regression parameters. Obviously, using GEE machinery, we can also assume different variance and covariance structure built in to the procedure. The difficulty is the justification of the structure chosen. They also illustrate how to fit Weibull regression models and piecewise exponential models by changing the offset or augmenting with a time-dependent covariate respectively.

Adjusting inference by design effects

In randomized clinical trials, the usual primary research question is what is the treatment difference among all the treatments? Let's assume that correlated observations form a cluster which can be a patient, a family or a community, etc., and assume the observations between clusters are independent. The idea behind design effect approach we are using is first to derive Taylor linearization for implicitly defined parameter vectors, which was developed by Binder (1983) in generalized linear models, and then apply a between-cluster variance estimator for the linearized statistic, as described by Bieler and Williams (1995). The details of the design effect approach for our case are the following. Let m_i be the number of clusters randomized to the treatment *i*, $i = 1, 2, \ldots, G$; n_{ij} be the number of observations for cluster *j* in *i*th treatment, $j = 1, 2, \ldots, m_i$; δ_{ijk} be the censoring indicator from the *k*th observation of *j*th cluster from *i*th treatment, k =1, 2,…, n_{ij};

$$\delta_{ij} = \sum_{k=1}^{n_{ij}} \delta_{ijk};$$

$x_{ijk} = (x_{1,ijk}, x_{2,ijk}, ..., x_{p,ijk})'$ be the vector of covariates (i.e. treatment, sex, race, etc.) for the *ijk*th observation; $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ and $E(\delta_{ijk}) = \mu_{ijk}$. From earlier developments (section 2), treat δ_{ijk} as if it were Poisson, a log link function is used in the generalized linear model, i.e. $\log \mu_{ijk} = x'_{ijk}\beta$ and

$$p(\delta_{ijk} \mid x_{ijk}, \beta) = e^{-\mu_{ijk}} (\mu_{ijk})^{\delta_{ijk}} / \delta_{ijk}!.$$

Thus,

$$\begin{aligned}\log(p(\delta_{ijk} \mid x_{ijk}, \beta)) &\propto -\mu_{ijk} + \delta_{ijk} \log(\mu_{ijk}) \\ &= -\Lambda_0(t_{ijk}) \exp(x'_{ijk}\beta) \\ &+ \delta_{ijk}(\log(\Lambda_0(t_{ijk})) + x'_{ijk}\beta).\end{aligned}$$

The log-likelihood equations are then

$$\begin{aligned}l(\beta) &= -\sum_i \sum_j \sum_k \Lambda_0(t_{ijk}) \exp(x'_{ijk}\beta) \\ &+ \sum_i \sum_j \sum_k \delta_{ijk}(\log(\Lambda_0(t_{ijk})) + x'_{ijk}\beta).\end{aligned}$$

The set of estimated Poisson regression coefficients, $\hat{\beta}$, that maximize $l(\beta)$ are found by solving the following score equations:

$$\begin{aligned}U(\beta) \equiv \frac{\partial l(\beta)}{\partial \beta} &= -\sum_i \sum_j \sum_k \Lambda_0(t_{ijk}) \exp(x'_{ijk}\beta) x_{ijk} \\ &+ \sum_i \sum_j \sum_k \delta_{ijk} x_{ijk} = 0.\end{aligned}$$

This can be done using the Newton-Raphson method. Then

$$\hat{\mu}_{ijk} = \Lambda_0(t_{ijk})\exp(x'_{ijk}\,\hat{\beta}).$$

Since μ_{ijk} may contain other nuisance parameters, we have to estimate them from the likelihood function. For example, if we assume Weibull baseline hazard, $\lambda_0(t) = \nu t^{\nu-1}$, the shape parameter ν can be estimated by

$$\hat{\nu} = \frac{\sum \delta_{ijk}}{\sum \log(t_{ijk})(\hat{\mu}_{ijk} - \delta_{ijk})},$$

and an iterative procedure can be used to find the estimates of β and other nuisance parameters.

The associated sample information matrix for estimating β is

$$I = -\frac{\partial^2 l(\hat{\beta})}{\partial\hat{\beta}\hat{\beta}'}$$
$$= -\sum_i\sum_j\sum_k \Lambda_0(t_{ijk})\exp(x'_{ijk}\hat{\beta})x_{ijk}x'_{ijk} + 0$$
$$= -\sum_i\sum_j\sum_k \hat{\mu}_{ijk}x_{ijk}x'_{ijk}. \tag{2}$$

Under cluster sampling, the inverse of the information matrix is no longer a valid estimate of the variance $\hat{\beta}$ (Binder, 1983). To address this problem, Binder (1983) gave a general method for deriving the variance of parameter estimators under clustering in survey sampling, which satisfy estimating equations of the form:

$$U(\hat{\beta}) = \sum_l u_l(\hat{\beta}),$$

where the sum is over the observations. Thus, using Taylor series linearization:

$$U(\hat{\beta}) = U(E(\hat{\beta}))$$
$$+\frac{\partial U(E(\hat{\beta}))}{\partial\hat{\beta}}(\hat{\beta} - E(\hat{\beta})) + o(\hat{\beta} - E(\hat{\beta})).$$

By the delta method, the variance of $\hat{\beta}$ is then estimated by

$$\hat{V}(\hat{\beta}) = (I^{-1})V_U(I^{-1})', \tag{3}$$

where

$$V_U = \hat{V}[U(\hat{\beta})].$$

Binder (1983) gave conditions under which (3) consistently estimates the asymptotic variance of $\hat{\beta}$. In order to obtain a cluster covariance matrix of $U(\hat{\beta})$, we first linearize $U(\hat{\beta})$, and then apply a between-cluster variance estimator for the linearized statistic. To this end, let

$$Z_{ijk} = x'_{ijk}\,\hat{r}_{ijk}$$

where $\hat{r}_{ijk} = \delta_{ijk} - \mu_{ijk}$ is the residual for the *k*th observation of the *j*th cluster from the *i*th treatment group. Accumulations of these linearized vectors are first formed at the cluster level, namely,

$$Z_{ij} = \sum_k Z_{ijk}. \tag{4}$$

The associated between-cluster within treatment group mean square matrix is

$$S_z = \sum_i m_i S_{zi},$$

where m_i denotes the number of clusters in treatment group *i* and

$$S_{zi} = \sum_j (Z_{ij} - \overline{Z}_i)(Z_{ij} - \overline{Z}_i)'/(m_i - 1),$$

depicts the *p x p* matrix of sample mean squares and cross products from treatment group *i*, with

$$\overline{Z}_i = \sum_j \frac{Z_{ij}}{m_i}.$$

Following (3), the estimated variance for β is given by

$$\hat{V}(\beta) = I^{-1} S_z (I^{-1})'.$$

The above estimate of the variance of $\hat{\beta}$ is called a "modified sandwich estimate" and converges to the true variance of $\hat{\beta}$ when the number of the clusters tends to infinity (Binder, 1983). If the total number of the clusters is small, then this estimate will be sharply biased towards zero, and some other estimate must be considered. Generally speaking, when the clusters are independent, the sum of the linearized vectors for each cluster, Z_{ij} in (4) can be unbiasedly estimated because $\hat{\beta}$ is usually a consistent estimate of β under usual regularity conditions without taking the correlation structure into account. Unlike the quasi-likelihood GEE approach of Liang and Zeger (1986), explicit specification of a correlation structure in the cluster is unnecessary, which is also mentioned in Bieler and Williams (1995).

Methodology

Generally speaking, there are two approaches for analyzing multivariate survival data. One is conditional model, and other is a marginal model. Conditional models induce dependence by including frailties (random effects) while marginal approach directly models fixed effects. We will employ a marginal-based approach when conducting simulations in order to evaluate the performance of the proposed design effect based approach. We specify a marginal survival distribution, and estimate the parameters characterizing the distribution. This approach however does not define the joint distributions for generating multivariate survival data, and thus the effect of dependence in repeating events over time cannot be studied. Hence we use a random effects approach as in Segal and Neuhaus (1993) where the joint distributions are forced to have proportional margins and a patterned covariance matrix.

We use positive stable mixing distributions (Hougaard, 1986) along with the random effects approach. Let T_{ijk} be the survival times of observation k of individual j with treatment group I conditional on an observed covariate ζ_j. In this setup we assume that T_{ijk} 's in different clusters are independent.

Now assume ζ to be positive stable with index α. The Laplace transform for ζ is $E(\exp(-s\zeta)) = \exp(-s^{\alpha})$. If we now define another random variable Y_{ijk} to be Weibull distributed with scale parameter $\exp(\beta' x_{ijk})$ and shape parameter a, then $T_{ijk} = Y_{ijk} \zeta_j^{-1/a}$.

Thus, the T_{ijk}'s within a cluster are multivariate Weibull with Weibull margins having scale $\exp(\alpha\beta' x_{ij})$ and shape αa. The correlation between $\log(T_{ijk})$ and $\log(T_{ijl})$ is then just $1-\alpha^2$ for $k \neq l$. The generation of positive stable variates ζ_j can be done using Splus which employs Chambers et al.'s (1976) algorithm.

In order to examine the performance of the newly proposed method for estimating regression parameters, we studied a number of scenarios. We first looked at varying the cluster size from $k = 5,10$ and also the number of clusters $C = 20,50$. The survival data was generated using the procedure just described with shape parameter $\alpha = 2$ and one covariate $\beta = 3$ for simplicity which are chosen arbitrarily.

The index of the positive stable distribution α was varied from 0.3 to 0.7 indicating decreasing levels of correlation between log survival times within a cluster. Survival times were censored at fixed times instead of random censoring to 10% and 20% censoring percentage. For each combination of experimental conditions, we conducted 200 simulations, and report biases of the regression parameter estimates from Poisson regression and GEE as well as mean variance of three types, i.e. naive, robust and new approach.

We fit Segal and Neuhaus's GEE-based method with independence correlation structures and compared the performance to the new method. The comparison will be made in terms of bias and variance. Since there is no explicit

formula available for the variance of $\hat{\beta}$ in this complex situation, so we don't know the true variance of $\hat{\beta}$. We use following approach to check the underestimation or overestimation of the estimate from each method in this finite sample situation. Let B be the number of simulations (in our case, we set $B = 200$, β_i, $i = 1,..., p$ be the true value of the coefficients, $\hat{\beta}_{ij}$ be the estimates of $\hat{\beta}_i$ in iteration j, where $i = 1,..., B$, and $\hat{\sigma}^2_{i,j}$ be the variance estimate of $\hat{\beta}_i$ in jth simulation after correction which accounts for the correlation of survival times within each cluster, then one way to check the biases of the variance estimate is the following efficiency quantity:

$$r_i = \frac{\hat{\bar{\sigma}}_i^2}{(m(\tilde{\beta}_i))^2 + \mathrm{var}(\tilde{\beta}_i)},$$

where

$$\hat{\bar{\sigma}}_i^2 = \sum_j \hat{\sigma}_{ij}^2 / B,$$

$$\tilde{\beta}_i = (\hat{\beta}_{i,1} - \beta_i, ..., \hat{\beta}_{i,B} - \beta_i)'$$

and $m(\tilde{\beta}_i)$ is the sample mean of $\tilde{\beta}_i$ and $\mathrm{var}(\tilde{\beta}_i)$ is the sample variance of $\tilde{\beta}_i$. If $r_i > 1$, then the variance is empirically overestimated, if $r_i < 1$, then the variance is empirically underestimated.

The simulation was conducted in *S-plus*. Our approach can be implemented with minor programming, a call of glm function and several other lines of coding for matrix manipulation (the program is available upon request).

Tables 1-4 give the results of the simulation. Notice first, as number of clusters increases, the smaller the bias in estimating the scale parameter *a*, and the regression coefficient β for Segal and Neuhaus's approach and our approach.

This is because the estimates are consistent when the number of clusters gets large; and there is no systematic difference of the biases when the cluster size, percentage of censoring, and value of index parameter α change.

Secondly, the variance estimate of $\hat{\beta}$ by the new method, the robust variance as well as naive variance estimates decrease when the number of cluster increases. Varying the cluster size does not change the variance, and there is no obvious evidence that a different percentage of censoring gives substantially different results. But increasing value of the index α, which changes the correlation of survival times in each cluster, does decrease the variance estimate in all three different types of estimates. This is because increasing α decreases correlation among the survival times within each cluster.

The naive variance estimates overestimate or underestimate the variance badly; the robust variance estimate and the new method usually underestimate the variance except in one case by our method with $r = 1.008$ ($C = 20$, $cen = 20\%$ and $\alpha = 0.4$). Overall, our method gives r values closer to 1 than the GEE approach, because correlation structure is not needed explicitly in calculating the variance of $\hat{\beta}$ as it is in GEE approach. The larger the number of clusters is the closer the r values are to 1.

A real data example (CGD)

The well-known Chronic Granulomatous Disease (CGD) dataset, which is described in the Appendix D of Fleming and Harrington (1991), has been analyzed by many authors. CGD is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections which usually present early in life and may lead to death in childhood. Phagocytes from CGD patients ingest microorganisms normally but fail to kill them, primarily due to the inability to generate a respiratory burst dependent on the production of superoxide and other toxic oxygen metabolites.

Thus, it is the failure to generate microbicidal oxygen metabolites within the phagocytes of CGD patients. There is evidence that gamma interferon is an important macrophage activating factor which could restore superoxide anion production and bacterial killing by phagocytes in CGD patients.

Table 1: Results for simulated multivariate Weibull distribution with number of clusters = 20 and 10% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.
In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and cens is percentage of censoring.

M_i = 20, *cens* = 10%	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7
Bias of a, scale parameter					
k = 5	0.1052	0.1605	0.1236	0.1449	0.09872
k = 10	0.1429	0.1195	0.1180	0.1118	0.09502
bias of b, Poisson					
k = 5	0.09133	0.06826	0.12269	0.21668	0.1123
k = 10	0.15252	0.11607	0.06024	-0.01395	0.1347
bias of b, GEE					
k = 5	0.09138	0.06829	0.12272	0.21671	0.1123
k = 10	0.15256	0.11610	0.06027	-0.01392	0.1347
variance of b, mod. rob.					
k = 5	2.674	1.364	0.7706	0.4536	0.2924
k = 10	2.546	1.311	0.7452	0.4266	0.2410
variance of b, naive					
k = 5	3.517	1.6908	1.4923	0.8766	0.6618
k = 10	1.511	0.8239	0.7739	0.4706	0.3094
variance of b, robust					
k = 5	2.356	1.201	0.6781	0.3993	0.2575
k = 10	2.243	1.155	0.6558	0.3754	0.2121
efficiency (r), new app.					
k = 5	0.8457	0.9160	0.7519	0.6152	0.7793
k = 10	0.7825	0.7471	0.8183	0.8497	0.6449
efficiency (r), naive					
k = 5	1.1124	1.1352	1.4561	1.1891	1.7641
k = 10	0.4643	0.4694	0.8498	0.9373	0.8281
efficiency (r), robust					
k = 5	0.7451	0.8066	0.6617	0.5416	0.6863
k = 10	0.6893	0.6580	0.7202	0.7477	0.5676

In order to study the ability of gamma interferon to reduce the rate of serious infections, a double-blinded clinical trial was conducted in which patients were randomized to placebo vs. gamma interferon. The data used here, which is a little different from that was used by Fleming and Harrington in the example (on page 162), has 65 patients in the placebo group, 63 in gamma interferon group, of 30 placebo patients who experienced at least one infection, 4 experienced 2, 4 experienced 3, 1 experienced 4, 1 experienced 5 and 1 experienced 7; of 14 treatment patients who experienced at least one infection, 4 experienced 2 and 1 experienced 3.

Table 2: Results for simulated multivariate Weibull distribution with number of clusters = 20 and 20% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.
In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and cens is percentage of censoring.

M_i = 20, *cens* = 20%	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7
Bias of a, scale parameter					
k = 5	0.1057	0.1744	0.1304	0.1441	0.10299
k = 10	0.1357	0.1215	0.1276	0.1165	0.09683
bias of b, Poisson					
k = 5	0.1076	0.04897	0.10081	0.19289	0.08116
k = 10	0.1658	0.10761	0.02919	-0.05122	0.11866
bias of b, GEE					
k = 5	0.1078	0.04907	0.10092	0.19301	0.08126
k = 10	0.1659	0.10773	0.02929	-0.05112	0.11877
variance of b, mod. rob.					
k = 5	3.308	1.690	1.0056	0.6356	0.4219
k = 10	3.120	1.642	0.9672	0.5725	0.3428
variance of b, naive					
k = 5	4.064	2.159	1.8301	1.1675	0.919
k = 10	1.809	1.011	0.9245	0.6216	0.433
variance of b, robust					
k = 5	2.921	1.493	0.8879	0.5616	0.3725
k = 10	2.751	1.451	0.8540	0.5050	0.3028
efficiency (r), new app.					
k = 5	0.8984	1.008	0.9238	0.7869	0.9634
k = 10	0.8430	0.811	0.9650	0.9482	0.8272
efficiency (r), naive					
k = 5	1.1033	1.2876	1.6810	1.445	2.098
k = 10	0.4887	0.4993	0.9224	1.029	1.045
efficiency (r), robust					
k = 5	0.7931	0.8903	0.8156	0.6953	0.8506
k = 10	0.7432	0.7164	0.8520	0.8364	0.7306

Table 3: Results for simulated multivariate Weibull distribution with number of clusters = 50 and 10% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.
In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and cens is percentage of censoring.

M_i = 50, *cens* = 10%	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7
Bias of a, scale parameter					
k = 5	0.04065	0.04009	0.06156	0.05327	0.04993
k = 10	0.05702	0.05781	0.07181	0.04742	0.02708
bias of b, Poisson					
k = 5	-0.01885	0.01177	0.03810	0.02688	0.04290
k = 10	0.10574	0.04004	0.09806	0.12305	0.02239
bias of b, GEE					
k = 5	-0.01884	0.01178	0.03810	0.02689	0.04291
k = 10	0.10575	0.04005	0.09806	0.12306	0.02240
variance of b, mod. rob.					
k = 5	0.9105	0.4661	0.2606	0.1546	0.09552
k = 10	0.8890	0.4462	0.2436	0.1422	0.08358
variance of b, naive					
k = 5	1.2961	0.7901	0.5114	0.4461	0.2702
k = 10	0.8169	0.5408	0.2750	0.2409	0.1648
variance of b, robust					
k = 5	0.8738	0.4476	0.2501	0.1484	0.09171
k = 10	0.8532	0.4285	0.2338	0.1365	0.08022
efficiency (r), new app.					
k = 5	0.860	0.9513	0.8769	0.6957	0.7103
k = 10	0.955	0.8215	0.7185	0.8391	0.6254
efficiency (r), naive					
k = 5	1.2242	1.6126	1.7212	2.008	2.009
k = 10	0.8775	0.9956	0.8111	1.422	1.233
efficiency (r), robust					
k = 5	0.8253	0.9135	0.8418	0.6678	0.6819
k = 10	0.9165	0.7889	0.6896	0.8056	0.6003

Table 4: Results for simulated multivariate Weibull distribution with number of clusters = 50 and 20% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.
In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and cens is percentage of censoring.

M_i = 50, *cens* = 20%	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7
Bias of a, scale parameter					
k = 5	0.04621	0.04515	0.0594	0.04871	0.05098
k = 10	0.05784	0.06741	0.0750	0.05162	0.02529
bias of b, Poisson					
k = 5	-0.02937	-0.0003165	0.03197	0.02579	0.03369
k = 10	0.11790	0.0056062	0.08744	0.11434	0.01731
bias of b, GEE					
k = 5	-0.02928	-0.0002196	0.03208	0.02589	0.03379
k = 10	0.11803	0.00057196	0.08756	0.11443	0.01741
variance of b, mod. rob.					
k = 5	1.064	0.5632	0.3231	0.1986	0.1235
k = 10	1.040	0.5327	0.3021	0.1795	0.1089
variance of b, naive					
k = 5	1.5115	0.9028	0.6095	0.5428	0.3484
k = 10	0.9288	0.6249	0.3331	0.2876	0.2029
variance of b, robust					
k = 5	1.027	0.5440	0.3129	0.1926	0.1198
k = 10	1.005	0.5153	0.2930	0.1739	0.1057
efficiency (r), new app.					
k = 5	0.8992	0.9887	0.9812	0.8340	0.8235
k = 10	0.9821	0.8901	0.8226	0.9614	0.7481
efficiency (r), naive					
k = 5	1.2774	1.585	1.8508	2.28	2.323
k = 10	0.8765	1.044	0.9069	1.54	1.393
efficiency (r), robust					
k = 5	0.8676	0.9551	0.9502	0.8088	0.7987
k = 10	0.9490	0.8609	0.7976	0.9311	0.7259

In order to check how our method works in the real data situation, we fit the CGD using the newly proposed approach with single treatment indicator covariate without controlling other covariates. As we can see in Table 5, the coefficients from our method and Segal and Neuhaus's method with independent working correlation structure are the same, and the coefficients using Andersen-Gill's and Cox model are similar. In Cox model, only the first event was used. The former (our method and Segal and Neuhaus's) is different from the latter (Andersen-Gill's model and Cox model) because the models are different; the coefficients are proportional by a constant, which is the index parameter in the positive stable distribution. Currently, to obtain an estimate of this correlation parameter is problematic as mentioned in Segal and Neuhaus (1993). Nevertheless, the ratio of $\hat{\beta} / s.e(\hat{\beta})$ from our method is comparable with that from Andersen-Gill model. Thus, our method is effective to detect significance of the treatment effect (gamma interferon) though the coefficient is underestimated since the index from the positive stable distribution is between 0 and 1.

Table 5: Results of fitting the CGD (Chronic Granulomatous Disease) dataset of various methods under consideration.

	$\hat{\beta}$	$s.e(\hat{\beta})$	$\lvert\hat{\beta}\rvert / s.e(\hat{\beta})$
New method	-0.856	0.2501	3.4226
Segal and Neuhaus	-0.856	0.2489	3.4389
Andersen-Gill	-1.2765	0.3774	3.3824
Cox model	-1.2062	0.4398	2.7426

Conclusion

It has been known that AG, WLW and PWP methods are extensions of survival models based on the Cox proportional hazards approach. They work well in one situation, but may not be appropriate in another (see Kelly and Lim, 2000, Therneau and Hamilton, 1997), since each method has different risk sets and risk intervals. Our new method was developed using a design effect approach from survey sampling and works well for the multivariate failure data. In addition, it's easy to implement. The strong assumption of the parametric form of the survival time can be relaxed by extending our method to the piecewise exponential case, which makes our method more flexible (Aitkin et. al., 1983). No covariance structure between the survival times in a cluster needs to be specified since it's implicitly built in our method.

As seen in our simulation study, the newly proposed method has slightly better finite sample performance than GEE based method. One limitation of our design effect method is that no time-dependent covariates are allowed. We also need to find a method to obtain an estimate of correlation parameter, as we saw it in Table 5; alternatively, a possible estimation strategy proposed by Segal, Neuhaus and James (1997) can be used for that. However, this limitation does not affect our ability to do inference about the regression parameters.

In our simulation, the censoring indicator is generated by fixed censoring time, a work on more general censoring mechanism, such as "independent censoring", is needed. In conclusion, the method of applying the cluster sampling techniques in the multiple failure data is a competing method of recent developed marginal approaches in the literature.

References

Aalen O. O. (1988). Heterogeneity in survival analysis, *Statistics in Medicine 7*, 1121-1137.

Aitkin M., & Clayton D. G. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics 29*, 156-163.

Aitkin M., Laird N., & Francis B., (1983). Reanalysis of the Stanford heart transplant data. *Journal of the American Statistical Association 78*, 264-274 (1983).

Andersen P. K., & Gill R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics 10*, 1100-1120.

Bieler G. S., & Williams R. L. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics 51*, 764-776.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review 51*, 279-292.

Chambers J. M., & Stuck B. W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association 71*, 340-344.

Clayton D. G. (1983). Fitting a general family of failure-time distributions using GLIM. *Applied Statistics 32*, 102-109.

Cochran, W. G. (1977). *Sampling Techniques*, Wiley, New York.

Fleming, T. R., & Harrington D. P. (1991). *Counting Processes And Survival Analysis*, John Wiley & Sons, New York.

Therneau, T. M., & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Hougaard P. (1986). A class of multivariate failure time distributions. *Biometrika 73*, 671-678.

Hougaard P. (1986). Survival models for heterogeneous population derived from stabe distributions. *Biometrika 73*, 387-396.

Huang X., Chen S., & Soong S. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics 54*, 1420-1433.

Kelly P. J., & Lim L. L-.Y. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine 19*, 13-33.

Laird N., & Oliver D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association 76*, 3231-240.

Liang K. Y., & Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13-22.

Lin D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine 13*, 2233-2247.

McGilchrist C. A., & Aisbett C. W. (1991). Regression with frailty in survival analysis. *Biometrics 47*, 461-466.

Norton K. I. et al,. (2001). Chronic radiographic lung changes in children with vertically transmitted HIV-I infection. *American Journal of Roentgenology 176*, 1553-1558.

Paik, M. C. (1988). Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics-Simulation 17*, 1155-1171.

Prentice R. L., Williams B. J., & Peterson A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika 68*, 373-379.

Rao, J., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics 48*, 577-585.

Rao, J., & Scott, A. J. (1999). A simple method for analysing overdispersion in clustered Poisson data. *Statistics in Medicine 48*, 577-585.

Segal M. R., & Neuhaus J. M. (1993). Robust inference for multivariate survival data. *Statistics in Medicine 12*, 1019-1031.

Segal M. R., Neuhaus J. M., & James I. R. (1997). Dependence estimation for marginal models of multivariate survival data. *Lifetime Data Analysis 3*, 251-268.

Therneau T. M., & Hamilton S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine 16*, 2029-2047.

Wei L. J., Lin, D.Y., & Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association 84*, 1065-1073.

Woodruff R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association 66*, 411-414.