The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 14

Marginal Models for Censored Longitudinal Cost Data: Appropriate Working Variance Matrices in Inverse-Probability-Weighted GEEs Can Improve Precision

Eleanor M. Pullenayegum, *McMaster University*
Andrew R. Willan, *Hospital for Sick Children*

Recommended Citation:

Pullenayegum, Eleanor M. and Willan, Andrew R. (2011) "Marginal Models for Censored Longitudinal Cost Data: Appropriate Working Variance Matrices in Inverse-Probability-Weighted GEEs Can Improve Precision," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 14.

DOI: 10.2202/1557-4679.1170

Marginal Models for Censored Longitudinal Cost Data: Appropriate Working Variance Matrices in Inverse-Probability-Weighted GEEs Can Improve Precision

Eleanor M. Pullenayegum and Andrew R. Willan

Abstract

When cost data are collected in a clinical study, interest centers on the between-treatment difference in mean cost. When censoring is present, the resulting loss of information can be limited by collecting cost data for several pre-specified time intervals, leading to censored longitudinal cost data. Most models for marginal costs stratify by time interval. However, in few other areas of biostatistics would we stratify by default. We argue that there are benefits to considering more general models: for example, in some settings, pooling regression coefficients across intervals can improve the precision of the estimated between-treatment difference in mean cost. Previous work has used inverse-probability-weighted GEEs coupled with an independent working variance to estimate parameters from these more general models. We show that the greatest precision benefits of non-stratified models are achieved by using more sophisticated working variance matrices.

KEYWORDS: censoring, longitudinal data, marginal mean model, inverse-probability weighting

Author Notes: The authors are funded through the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada; however, the opinions expressed are those of the authors and should not be attributed to any funding agency.

1. Introduction

Clinical trials are usually powered to detect efficacy or effectiveness, but not cost-effectiveness. Cost data are usually highly variable, and so the confidence intervals around estimates of cost-effectiveness will often be unhelpfully wide. It is perhaps this that has prompted much work on the efficiency of cost estimation (Bang and Tsiatis, 2000; Zhao *et al.*, 2007). There are three important factors impacting the precision of estimates: the amount of information lost to censoring, the efficiency of the estimator, and the choice of model. The second of these factors has received much attention, but the third has not.

It is usually not easy to alter the censoring fraction; however the amount of cost information lost to censoring can be reduced by collecting cost information over narrower time intervals. Note that an individual contributes cost information for all intervals up to the interval in which censoring occurs. When cost is collected over several time intervals rather than just at the end of the study, those individuals who are lost to follow-up still contribute some cost information. Bang & Tsiatis (2000) describe the use of a partitioned inverse-probability weighted (IPW) estimator to use the additional cost information.

In the same paper, Bang & Tsiatis discuss improved efficiency estimators. These improved estimators use the semi-parametric efficiency results of Robins *et al.* to recover some of the information lost to censoring, and are available for both the simple and the partitioned Bang and Tsiatis (BT) estimators. However, the sheer complexity of these improved estimators, together with the lack of readily available computer code, make their use impractical. Indeed, Bang and Tsiatis' paper on IPW and improved estimators of mean cost has been much cited (we found 64 citations in Web of Science by July 2008); however to the best of our knowledge (we were unable to retrieve the full text of three of the citing articles), none of the authors citing the paper used the improved efficiency estimators. When efficiency was a concern, the partitioned estimator was used.

We now turn to the impact of the model on the precision of cost estimates. The Bang and Tsiatis (BT) partitioned cost estimator and their improved estimator modelled each time interval separately, that is, the between-treatment difference in mean cost was estimated separately for each time interval and then summed to estimate the total between-treatment difference in mean cost. Although this is a valid approach, it is different from the model usually used for longitudinal data. More typical is a marginal mean model for the vector of cost intervals and parameter estimation using a GEE. These non-stratified models introduce the scope for pooling information across time intervals, *i.e.* for using the partitioned, longitudinal cost information, but without stratifying the model by time interval. This has the potential to improve precision, particularly when censoring is heavy in later intervals. Lin (2003) discusses GEE-type models for longitudinal cost data

subject to censoring, using IPW to maintain consistency, and allowing for non-stratified models. Lin takes the working correlation matrix in the GEE models to be the identity.

We note that a suitable choice of working variance matrix in Lin's GEE equations may be a simple way to improve efficiency, and one which could be implemented easily using existing software. Also, we note that Lin's GEE-type models have never been used in conjunction with efficient estimators proposed by Robins *et al.* (1995a)(1995b). In this paper, we give the improved efficiency version of Lin's IPW GEE estimator, which is extremely complex, then propose a less efficient, but more practical, simplified version. In Section 3 we demonstrate through a simulation study that our simplified estimator offers gains in precision over Lin's IPW GEE estimator, without becoming unmanageably complex. This improvement in precision is further demonstrated through use with a real dataset in Section 4. We close with some practical recommendations in Section 5.

2. Estimators of Between-Treatment Difference in Mean Cost

Notation

Suppose there are n individuals, and that the i^{th} individual has survival time T_i and censoring time U_i . Let $X_i = \min(T_i, U_i)$, $\delta_i = I(T_i \leq U_i)$, and $\bar{\delta}_i = I(U_i < T_i)$. If τ is the maximum follow-up time (*i.e.* the duration of interest), set $X_i^* = \min(X_i, \tau)$. Let G be the censoring survivor function, *i.e.* $G(t) = P(U_i \geq t)$. G can be estimated using a Kaplan-Meier estimator, or, in the case that G depends on covariates, through a proportional hazards regression combined with the Breslow estimate of the baseline hazard (see Pullenayegum and Willan (2007) for an example of how dependent censoring can be handled). Let $C_i(t)$ be the cost accrued by individual i by time t , and let $0=a_0, a_1, \dots, a_K=\tau$ be the time points at which patient follow-up occurs and cost information is collected. Let C_{ik} be the cost incurred by individual i in time interval k defined by $(a_{k-1}, a_k]$, (*i.e.* $C_{ik} = C_i(a_k) - C_i(a_{k-1})$), let C_{i+} be the total cost accrued by individual i from time 0 to τ , and let C_i be the K -vector of interval-specific costs for individual i . Let $X_{ik}^* = \min(X_i, a_k, \tau)$, and let $\delta_{ik}^* = 1$ if C_{ik} is observed, *i.e.* $\delta_{ik}^* = \delta_i + \bar{\delta}_i I(U_i \geq X_{ik}^*)$.

Allowing for time-dependent covariates, let Z_i be a $K \times p$ dimensional matrix whose k - l^{th} element Z_{ikl} is equal to the value of covariate l ($l = 1, 2, \dots, p$) for individual i during interval k . We shall take Z_{ik} to denote the p -vector of covariates for individual i for interval k . If the l^{th} covariate is not time-dependent, then $Z_{ik_1l} = Z_{ik_2l}$ for all i, k_1 and k_2 . In most cases we will take $Z_{ik1} = 1$ for all i and k , to provide an intercept, and take Z_{ik2} to be an indicator variable for treatment group. We note that in the context of a randomized trial, it would be unwise to adjust for time-dependent covariates representing patient characteristics that

change over time, as these may mediate the effects of treatment. The purpose of allowing for time-dependent covariates in this context is primarily to allow treatment effects on cost to vary over time. For example, setting $Z_{i12}=I[\text{Patient } i \text{ randomized to trial treatment}]$, $Z_{ik2} = 0$ for $k > 1$, $Z_{ik3}=I[\text{Patient } i \text{ randomized to trial treatment}]$ for all k , would allow for a separate treatment effect on cost for the first interval as compared to subsequent intervals. This might be a desirable feature if the two treatments to be compared had different start-up costs, as in the case of a drug being compared to a surgical procedure. Throughout the manuscript, we will use i to index individuals, j and k to index time intervals, and l to index covariates.

Cost-effectiveness evaluations require an estimate of the between-treatment difference in mean costs, the between-treatment difference in mean effectiveness, and the variance-covariance matrix between the two. Effectiveness is often measured in terms of survival or quality adjusted life years. In this paper, our focus will be on the mean difference in the cost between groups; efficacy or effectiveness outcomes can be estimated separately. As is usual for cost-effectiveness evaluations, the perspective will be that of the healthcare payer, so that costs drop to zero once an individual has died.

Efficiency

Many estimators for the between-treatment difference in mean cost have been proposed (Bang and Tsiatis, 2000; Lin *et al.*, 1997; Lin 2000; Liu *et al.*, 2007), and O'Hagan (2004) and Zhao *et al.* (2007) discuss equivalencies between them. Some of these estimators use partitioned cost data (Bang and Tsiatis, 2000; Lin, 2000), some use techniques to improve efficiency (Bang & Tsiatis, 2000). The Bang & Tsiatis inverse probability estimator for the mean cost in a treatment group is

$$\frac{1}{n} \sum_{i=1}^n \frac{\delta_i C_{i+}}{\hat{G}(X_i)} \quad (2.1)$$

where $\hat{G}$ is the Kaplan-Meier estimate of G . Thus the estimator of the mean cost is the mean of the fully observed costs weighted by the inverse of the probability of being observed. The corresponding partitioned estimator is

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{\delta_{ik}^* C_{ik}}{\hat{G}(X_{ik}^*)} \quad (2.2)$$

that is, the mean cost is calculated separately for each time interval, then the mean costs are summed over time intervals. Improved versions of both the simple and partitioned estimators are given in Bang & Tsiatis (2000). Zhao *et al.* (2007) show that when all the censoring points occur at the boundaries of the time intervals, the BT improved simple estimator is equivalent to the BT partitioned estimator. This means that when the boundaries of the time intervals include all the censoring points, the amount of information recovered by the improved IPW estimator is exactly equal to the amount of information gained by using the partitioned estimator. In this situation the BT partitioned estimator is thus to be preferred over the improved IPW estimator because it will be no less efficient and is easier to calculate and understand. The BT partitioned improved estimator can further improve efficiency; however it has rarely been used. In previous work we extended the BT improved partitioned estimator to the more general semi-parametric model (*i.e.* with covariates in addition to treatment group), and noted that whilst the improved efficiency estimator does improve upon the IPW estimator, greater gains in precision can be achieved by careful modelling of covariates. Even in a randomised trial, adjusting for covariates can explain some of the residual variance and thus improve precision.

Pooling

On a similar note, it is plausible that adopting more refined models than the usual stratified model used in cost estimation might be a more important factor in improving precision than using a complex estimator. The semi-parametric cost models described above stratify by time interval by default, *i.e.* estimate a separate treatment effect for each time interval, and then sum over intervals. Particularly when there are a large number of time intervals, it may be possible to improve precision by examining whether stratification is necessary and revising the model accordingly. It may be unnecessary to estimate a separate treatment effect on cost for each time interval. At the simplest level, we might take the treatment effects on mean cost to be equal across all intervals. Alternatively, we might allow the mean difference for the first interval to differ from all the subsequent intervals, to allow for the possibility of an initial cost of treatment. The modelling can be very flexible -- see Section 4 for an example.

Lin (Lin, 2003) previously proposed a GEE-modelling framework in which $E(C_{ik} | Z_i) = \alpha' Z_{ik}$ (rather than the stratify-by-default model, which is $E(C_{ik} | Z_i) = \alpha_k' Z_{ik}$). If there is no censoring α can be estimated by solving the GEE equations

$$\sum_{i=1}^n Z_i' V_i^{-1} (C_i - Z_i \alpha) = 0$$

however, since some data is lost to censoring, Lin uses the inverse-probability weighted GEE (IPWGEE) equations

$$\sum_{i=1}^n Z_i' V_i^{-1} D_i (C_i - Z_i \alpha) = 0 \quad (2.3)$$

where D_i is the diagonal matrix with the k^{th} diagonal entry equal to $\delta_{ik}^* / \hat{G}(X_{ik}^*)$ and $\hat{G}$ is the Kaplan-Meier estimator of G . In Lin's application, the working correlation matrix is taken to be the identity, and under the assumption of equal variances for each time period, V cancels out of the equations. See Baser (2006) for an alternative model for longitudinal cost data.

We now give the format of the semi-parametric efficient estimator. The purpose of showing this equation is simply to demonstrate that the estimator is too complex to be useful in many practical settings. The efficient estimator of α solves

$$\sum_{i=1}^n \sum_{j=1}^K \sum_{k=1}^K Z_{ik}' h_{eff}(Z_{i0})_{kj} D_{ij}^{eff} (C_{ij} - Z_{ij}' \alpha) = 0 \quad (2.4)$$

where j and k index time intervals and

$$h_{eff}(Z_{i0}) = \left(V(Z_{i0}) - \int_0^{\tau} \frac{E(\text{var}(C_i | C_{iu}) | Z_{i0})}{G(u)^2} dG(u) \right)^{-1}$$

$$\text{with } V(Z_{i0}) = \text{var}(C_{ij} - Z_{ij}' \alpha | Z_{i0})$$

Z_{i0} = those components of Z that are known at baseline

$$C_{iu} = \{T_i I[T_i \leq u], Z_{ik}, C_i(a_k) \text{ with } k : a_k \leq u\}$$

= individual i 's uncensored data up to time u .

$$D_{ijk}^{eff} = \frac{\delta_{ij}^*}{\hat{G}(X_{ij}^*, w_{ikj})}, \text{ with}$$

$$w_{ikj}(u) = \frac{Z_{ik}' h_{eff}(Z_{i0})_{kj} E(C_{ij} - Z_{ij}' \alpha | C_{iu})}{G(u)}$$

and $\hat{G}(t; w_{ikj})$ is a Cox proportional hazards estimate of G evaluated at t , using time-dependent covariates w_{ikj} (see Robins *et al.*, 1994, 1995a, 1995b). Intuitively, h_{eff} acts as a working variance matrix that accounts for the inverse probability weighting, and w uses the expected value of the residuals (costs minus marginal means) given observed information to recover information lost to censoring. Calculating the conditional expectations and variances required for h_{eff} and w requires some strong assumptions to reduce the dimensionality of the history C_{iu} (e.g. we might assume that C_{ij} depends on C_{iu} , $u < a_{j-1}$, only through the total cost accrued by individual i at time u).

Clearly, the deterrent to using this estimator is its complexity. Not only is it algebraically complex, but it also requires, for each pair of indices k, j in $1, \dots, K$, fitting a Cox proportional hazards model with time-dependent covariates w_{kj} , and hence computation time will increase quadratically with K . In many cases K will be large, as costs may be collected monthly or quarterly for the duration of a multi-year study, so that estimation becomes prohibitively slow. Conceptually, there are two ways in which this efficient estimator improves upon the IPW estimator. Firstly, some of the information lost to censoring is recovered through the use of the time-dependent covariates w . Observe that $\sum \delta_{ij}^* w_{ikj}(X_{ij}^*)$ is the contribution to the simple IPW score function for individual i in interval j (see equation (2.3)). Moreover, equation (2.4) may be re-written in the form of a doubly robust score function as

$$\sum_{i=1}^n \sum_{j=1}^K \sum_{k=1}^K \left(\delta_{ik}^* w_{ijk}(X_{ik}^*) + \int_0^{a_k} w_{ijk}(u) dM_i(u) \right) = 0$$

where $dM_i(u) = (1 - \delta_i)I(X_i = u) - I(X_i \geq u)\lambda(u)du$ defines is the martingale increment for the censoring process, and $\lambda(u)$ is the hazard function for the censoring process. The second term in this score function incorporates information from those lost to censoring by taking the expected value of the uncensored contribution conditional on the observed information.

The second way that the semi-parametric efficient estimator improves upon the IPW estimator is by using more efficient weight matrix $h_{eff}(\cdot)$ in place of the identity matrix. Observe that the weight matrix $h_{eff}(\cdot)$ does a similar job to the working correlation matrix in a GEE: when specified correctly it improves efficiency; however, the estimator is consistent when the working correlation matrix is misspecified. The GEE equations are asymptotically efficient when the working variance matrix is correctly specified, and it is easy to see that $h_{eff}(\cdot)$ reduces to the inverse of the true variance matrix in the absence of censoring. Thus $h_{eff}(\cdot)$ is the inverse variance matrix with an adjustment to account for the probability weights.

The key point of this paper is that it should be possible to compromise between the inefficient IPW estimates and the complex efficient estimates by making use of the weight matrix $h_{\text{eff}}(\cdot)$ without using the complex time-dependent covariates w . Define the simplified estimator as the solution to

$$\sum_{i=1}^n \sum_{j=1}^K \sum_{k=1}^K Z_{ik}' h_{\text{eff}}(Z_{i0})_{kj} D_{ij} (C_{ij} - Z_{ij}' \alpha) = 0 \quad (2.5)$$

where D_{ij} is once again the simple inverse-probability weight $\delta_{ij}^* / \hat{G}(X_{ij}^*)$, with $\hat{G}$ the simple Kaplan-Meier estimator of G .

The aim of this paper is to demonstrate that firstly there are gains in precision by choosing a non-stratified model, and secondly, the simplified improved estimator provides more precise estimates than either the BT partitioned or the working-independence IPWGEE estimator without being unmanageably complex.

3. Simulation Study

Simulation Set-up

For the simulations there were just two covariates: an intercept and a treatment group covariate, with half the individuals assigned to each group. Data were generated similarly to Bang & Tsiatis [2000], with

$$C_{ik} = I[k=1]c_i + (\min(T_i, a_k) - a_{k-1})_+ (d_i + e_{ik}) + I[a_{k-1} \leq T_i < a_k]d_i$$

where

$$\begin{pmatrix} \log(c_i) \\ \log(d_i) \end{pmatrix} \sim N \left(\begin{pmatrix} 9.170 \\ 7.32 \end{pmatrix}, \begin{pmatrix} 0.08 & 0.057 \\ 0.057 & 0.111 \end{pmatrix} \right)$$

e_{ik} were iid lognormal variables with mean 4.025 and variance σ^2 , the survival time T_i was given by

$$T_i = T_{0i} \exp(-0.25(\log(c_i) - \log(\bar{c})) - 0.3(\log(d_i) - \log(\bar{d}))),$$

with $\log(T_{0i}) \sim \text{iid } N(1.466, 0.288)$, and $\bar{c}$, $\bar{d}$ the sample means of the c_i and d_i respectively. The simulation set-up has an initial start-up cost, a cost around the time of death, and correlated costs for each intervening interval. Censoring was

uniform on (0, 20), resulting in 25% of individuals being censored. The time-span of interest was 0 to 10, the sample size was set to 50 per group, and each simulation was repeated 10,000 times. We looked at two values of K (2 and 5), with time intervals of equal lengths in each case. In order to assess the effect of autocorrelation in costs on the relative efficiency of the improved and simplified estimators, we considered two values of σ^2 : 0.435 and 4.35 to represent moderate and low autocorrelations respectively. Since the same data generation procedure was used for each treatment group, the true regression coefficient for the treatment group is zero for all time intervals.

We also investigated a non-null case where there was a difference in both costs and survival time between treatment groups. We set $K=2$, $\sigma^2=0.435$, and left all the parameters as before, except that the patients in the treatment group had

$$\begin{pmatrix} \log(c_i) \\ \log(d_i) \end{pmatrix} \sim N \left(\begin{pmatrix} 9.4794 \\ 6.32 \end{pmatrix}, \begin{pmatrix} 0.08 & 0.057 \\ 0.057 & 0.111 \end{pmatrix} \right)$$

These parameters were chosen so as to ensure that the treatment effect on cost was approximately equal for the two time intervals. Estimates of the correlations in costs across time intervals and the mean cost for each time interval, based on 1,000,000 simulated data points, are given in Table 1.

Five estimates of the between-treatment difference in cost were estimated for each simulated dataset. The first two were based on the completely stratified model $E(C_{ik}) = \alpha_{0k} + \alpha_{1k}I[group = 2]$, and used the BT partitioned estimator (equation (2.1)) and the BT improved partitioned estimator (equation (2.2)) respectively. The remaining estimates were based on a GEE-type model in which the intercept was different for each time period, but the treatment effect was constant ($E(C_{ik}) = \alpha_{0k} + \alpha_1I[group = 2]$). This model corresponds to pooling the treatment effects across time intervals. Under this model, we calculated Lin's IPWGEE estimator using equation (2.3) (which we shall call the pooled working independence estimator), the improved estimator (pooled improved, using an approximation to equation (2.4)), and the simplified improved estimator (pooled simplified, equation (2.5)). The parameter of interest was the between-group difference in mean total cost, *i.e.* $\Sigma\alpha_{ik}$ for the stratified model and $K\alpha_1$ for the pooled model.

Table 1: Estimated means and correlation matrices for the simulated cost data

<i>Means</i>	$E(C_{i1})$	$E(C_{i2})$	$E(C_{i3})$	$E(C_{i4})$	$E(C_{i5})$
K=2, moderate autocorrelation	17200	1700			
K=2, low autocorrelation	18900	2100			
K=5	13400	2900	1600	730	330
K=2, non-null: group1, group2 and difference between groups (gp2-gp1)	16611, 17130 Diff=518	929, 1439 Diff=510			
<i>Correlation matrices</i>					
K=2, moderate autocorrelation	$\begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$				
K=2, low autocorrelation	$\begin{pmatrix} 1 & 0.015 \\ 0.015 & 1 \end{pmatrix}$				
K=2, non-null: group1, group 2	$\begin{pmatrix} 1 & -0.001 \\ -0.001 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$				
K=5	$\begin{pmatrix} 1 & 0.3 & 0.004 & 0.07 & -0.08 \\ 0.3 & 1 & 0.3 & 0.1 & 0.07 \\ 0.004 & 0.3 & 1 & 0.6 & 0.2 \\ 0.07 & 0.1 & 0.6 & 1 & 0.6 \\ -0.08 & 0.07 & 0.2 & 0.6 & 1 \end{pmatrix}$				

The conditional moments required for the improved estimator of cost were calculated by inverse-probability weighted regression. The conditional expectation $E(C_{ik} | C_{iu})$ was estimated by a regression model, in which the observed costs C_{ik} , weighted by the inverse probability of their being observed, were regressed onto the individual's mean cost for time periods up to time u , *i.e.* if $\bar{C}_{iu}$ is the mean of the observed costs for individual i up to and including time u , then $E(C_{ik} | C_{iu})$ was estimated by regressing $\delta_i C_{ik} / \hat{G}(X_{ik}^*)$ onto $\bar{C}_{iu}$. Individuals who were censored or who had died before time u were excluded from the regression model for the conditional expectation, and it was assumed that once an individual had died there were no further costs. The regression approach was modified when estimating the conditional variances $\text{var}(C_i | C_{iu})$. For $k \geq j$ and $X_i^* > u$, the cross-product of the residual for cost at time k and the residual for cost at time j was calculated, weighted by the probability of the cost at time k being observed (given the individual was alive and under observation at time u). Each individual's mean cost up to time u was also calculated. Rather than

regressing the cross-product onto the mean cost itself, thus risking negative predicted variances, the individuals' mean costs were categorised into quintiles and the mean of the cross-products of the residuals within each quintile was found. $\text{var}(C_i|C_{iu})$ was assumed to be zero for individuals who died before time u and was otherwise assigned the average of the cross-product from the appropriate quintile of mean cost up to time u . That is, if $\delta_i=1$ and $X_i < u$, then $\text{var}(C_i|C_{iu})=0$, otherwise

$$\text{var}(C_{ij}, C_{ik} | C_{iu}) = E \left(\frac{\delta_{ik} (C_{ij} - \hat{E}(C_{ij} | \bar{C}_{iu})) (C_{ik} - \hat{E}(C_{ik} | \bar{C}_{iu}))}{G(X_{ik})} \middle| \text{Quintile}(\bar{C}_{iu}) \right)$$

Finally, the expected values of the conditional variances were estimated by taking treatment group means $E(\text{var}(C_i | C_{iu}) | Z_i)$.

In calculating these conditional moments given C_{iu} , we have made the assumption that $\bar{C}_{iu}$, coupled with survival status up to time u , captures all the important information in C_{iu} . This is a strong assumption that is unlikely to hold in general, however some such assumption is needed to reduce the dimensionality of C_{iu} . It is for this reason that implementations of the efficient estimators are referred to as “improved” rather than “efficient”: they are only approximations to the efficient estimators (Robins and Rotnitzky, 1992).

Simulation Results

Table 2 shows the observed standard errors amongst the 10,000 samples for each of the five estimators. Note that the biases are all small compared to the standard errors, and so the root mean square errors and the observed standard errors are very close. These results show that, under the conditions studied, there is little if any benefit to pooling if using a working independence correlation structure, that the improved estimators are more efficient than their IPW counterparts, that with more time intervals the degree of improvement from using a pooled model appears to increase, provided we do not assume working independence, and that the pooled simplified estimator is less efficient than the pooled improved estimator but more efficient than either the pooled working-independence estimator or the partitioned BT estimator.

The observed standard errors of the pooled working-independence and simplified estimators are shown in Table 3, alongside the average estimated standard errors, and the coverage of the estimated 95% confidence intervals derived from these estimated standard errors. The coverage probabilities for the pooled working-independence estimator are acceptable, however those for the

pooled simplified estimator are too small, likely because the estimated standard errors do not account for the variability of the estimated weight matrix h_{eff} .

These results suggest that, under the conditions studied, and provided that pooling is appropriate, the pooled simplified estimator will outperform both the BT partitioned estimator and Lin's pooled working-independence estimator. The size of the gain varies considerably, but appears to increase with increasing numbers of time intervals and decreasing autocorrelation in costs. Moreover, there appears to be little benefit to pooling if using the working-independence estimator. Since the variance estimator for the pooled simplified estimator is too small, we suggest using bootstrap methods to derive p-values and confidence intervals.

This simulation study represents a simple case where the between-treatment difference in costs is constant over time. In practice, this will not always hold, but the models can handle these more complex settings, as we now illustrate through an example.

4. Example: The Canadian Implantable Defibrillator Study

The Canadian Implantable Defibrillator Study (CIDS) was a randomised trial of 659 patients with resuscitated ventricular fibrillation or sustained ventricular tachycardia or with unmonitored syncope. Between October 1990 and January 1997, patients were randomised between amiodarone and an implantable cardioverter defibrillator. The primary outcome was all-cause mortality, and the clinical results can be found in Connolly *et al.* (2000). Due to budgetary constraints, costs were collected only for the first 430 patients, and the economic findings are reported in O'Brien *et al.* (2000). For the purposes of the economic analysis, the duration of interest is $\tau = 6.42$ years (77 months). Since not all patients were followed for this long, censoring is present. The time points for collection of cost data, expressed in days, were $(a_1, \dots, a_K) = (0, 914, 1005, 1096, 1187, 1279, 1370, 1461, 1553, 1644, 1735, 1827, 1918, 2009, 2100, 2192, 2238, 2310)$, with $K=17$. There was no significant difference between groups in mortality ($p=0.142$), with both groups experiencing approximately 30% mortality by the end of year 4.

Table 2: Observed standard errors (bias), for varying numbers of time intervals and varying degree of correlation between costs for the same individual across time intervals. The BT partitioned estimator is the usual Bang & Tsiatis partitioned estimator, and the partitioned improved estimator is the improved efficiency modification. The remaining estimators assume a common treatment effect across the time intervals and use inverse-probability weighted GEEs. The pooled working independence takes the working independence matrix to be the identity, while the pooled simplified estimator uses an approximation of the asymptotically efficient working variance matrix. The pooled improved estimator is an approximation to the asymptotically efficient estimator.

Estimator Simulation Parameters	BT partitioned	Partitioned improved	Pooled working independence	Pooled simplified	Pooled improved
K=5	1290 (-5.9)	1257 (-75.3)	1290 (-10.4)	1157 (-11.7)	1069 (-8.9)
K=2	1361 (-1.4)	1307 (-56.7)	1364 (-2.7)	1333 (-0.84)	1277 (2.1)
K=2 – less autocorrelation	4570 (-9.6)	4385 (-75.9)	4602 (-23.1)	1887 (1.19)	1897 (-1.63)
K=2, non-null	1180 (7.8)	1173 (13.4)	1178 (-29.1)	952 (14.9)	960 (7.0)

Table 3: Observed Standard Error, Average Estimated Standard Error and 95% confidence interval coverage rate for the pooled working-independence and simplified estimators

	Pooled working independence	Pooled simplified
K=5	1290, 1355, 95.4%	1157, 1036, 90.9%
K=2	1364, 1353, 94.2%	1333, 1244, 92.8%
K=2 – less autocorrelation	4602, 4860, 94.7%	1897, 1787, 93.0%
K=2, non-null	1178, 1136, 94.3%	952, 891, 91.8%

Conditional cost expectations given information up to and including time a_k , which are needed for the improved estimators, were calculated by an inverse-probability weighted regression of interval costs on the mean of the cost up to the k^{th} interval, but excluding the costs from the first 90 days, in recognition that the initial treatment costs may not be predictive of subsequent costs. Note that these conditional costs are used only in calculating the improved inverse probability weights; the between-treatment difference in costs uses cost data from the entire duration of the trial, including the first 90 days.

Initially, the usual stratified model was used, and we calculated the partitioned BT and the improved partitioned BT estimators. The estimates of between-treatment difference were similar for the two estimators, with an estimated increase of approximately \$49,000 for the defibrillator group as compared to the amiodarone group (Table 4). The improved estimator achieved a small reduction in the estimated standard error when compared to the partitioned BT estimator.

Modelling

Marginal Mean Model

A plot of the inverse-probability estimates of the treatment effects suggests that after the long initial interval, the treatment contrast for the first quarter of the year appears to be consistently higher than in the other quarters, and treatment contrasts in years 3 and 4 are generally higher than those nearer the beginning or end of the study (Figure 1). Taking this into account, the revised model includes treatment interactions with each of four dummy variables indicating year of study ($2 \leq k \leq 4$, $5 \leq k \leq 8$, $9 \leq k \leq 12$, $13 \leq k \leq 17$), and each of three dummy variables representing quarters 2, 3, and 4 relative to the first quarter. In addition, the intercept is modelled as completely time-varying. We shall call the revised model the year-and-quarter model:

$$E(C_{ik}) = \alpha_k + Z_i (I(k=1) + Year_k + Quarter_k)$$

where Z_i is a 0/1 treatment group indicator, Year is the factor corresponding to year of study and Quarter is the factor corresponding to quarter of study.

Working Correlation Model

The matrices $E(\text{var}(C_i | C_{iu}) | Z_{i0})$ and $\text{var}(C_i | Z_{i0})$ were assumed to be block diagonal, with a 1×1 block for the first time interval and a 16×16 block for the remaining intervals. We estimated the correlation in costs for times a_k to a_K , given data prior to a_k , using a method of moments (Prentice, 1988). We calculated the conditional residuals (costs at each time interval minus their conditional means given data up to a_k), and standardised so that the residuals for each group and each time point had unit variance. Noting that the covariances would be zero for those who had been observed to die, we restricted attention to those who were still under observation at time a_k and known to be alive. The within-individual pairwise cross-products of the residuals were calculated, weighted by the inverse probability of the cross-product being observed, then the correlation was modelled through regression (using GEEs to account for within-individual clustering in the cross-products). Specifically, the weighted cross-products were regressed onto the lag between the two residuals (as a categorical covariate) and the time-period corresponding to the earlier of the residuals in the cross-product, excluding lag zeros, so that the model for the correlation is $\text{corr}(C_{ik}, C_{ij} | \text{alive at } t_k) = a_j + b_{|k-j|}$. This structure was chosen because it required estimation of considerably fewer parameters than a completely unstructured matrix, whilst still being more general than an autoregressive or exchangeable structure. Moreover, the time period and lag were statistically significant for several of the conditional correlation matrices corresponding to $\text{var}(C_i | C_{iu})$, particularly for higher values of u . The expected values from this regression model were used to construct the correlation matrix, and hence the variance-covariance matrix amongst those known to have survived. This was weighted by the probability of surviving to give the variance-covariance matrix amongst the population as a whole.

Figure 1: Inverse-probability weighted estimates of the treatment effect for each of the time intervals, using the usual univariate model, which stratifies by time interval, and the year-and-quarter model.

Results

Using this model, we computed the working independence IPWGEE estimator and the simplified IPWGEE estimator (Table 4). The estimates are similar. The year-and-quarter working-independence estimator had a smaller standard error than that for the usual stratified model (3933 as compared to 3970). The simplified improved year-and-quarter estimator had the smallest standard error at 3726 (compared to 3943 for the improved stratified estimator). The standard error

for the simplified improved estimator is the bootstrap standard error (based on 1000 samples), as the theoretical standard error may be too small. For comparison, the bootstrap standard error for the working-independence IPWGEE estimator is 3953, *i.e.* similar to the theoretical standard error. Thus the simplified year-and-quarter estimator achieves a 12% reduction in variance over the traditionally used BT partitioned estimator. Whilst this may sound modest, achieving this level of reduction in variance by increasing the sample size would have required approximately 50 ($=430 \cdot (1/0.88 - 1)$) more patients.

Table 4: Estimates of between-treatment difference in cost and corresponding standard errors (SE) using the Bang & Tsiatis partitioned estimator, the partitioned improved estimator, the working-independence GEE estimator (with intercepts varying for each time period and treatment effect depending on year and quarter), and the simplified GEE estimator. Theoretical standard errors are not calculated for the simplified year-and-quarter model, as they are known to underestimate true variability.

Estimator	Between-treatment difference in mean costs	SE (t: theoretical; b: bootstrap)
BT partitioned	48728	3970 (t)
Partitioned improved	48300	3943 (t)
Working-independence Year-and-Quarter	48726	3933 (t); 3953 (b)
Simplified Year-and-Quarter	48518	3726 (b)

5. Discussion

In this article we have argued that, just as the partitioned Bang-Tsiatis estimator improves upon the unpartitioned BT estimator, there are gains in precision upon using a non-stratified model. Typically cost analyses stratify the treatment effect on costs by time interval by default, in contrast to effectiveness analyses where by default the treatment effect is usually assumed to be constant across time intervals. Lin (2003) first introduced GEE-type models for cost data, and estimated the parameters using inverse-probability weighting in conjunction with an independent working correlation structure; our results suggest that most of the gains in efficiency on using non-stratified models can be achieved by using a more general correlation structure.

The reason why pooling improves precision for the simplified estimator but not the working independence estimator is intuitive. The benefit of pooling is that it allows later time intervals subject to greater censoring to borrow strength

from earlier intervals with less censoring. This advantage is lost, however, if the independent working correlation matrix is used in conjunction with the assumption of equal variances across intervals, as then all intervals contribute equally, regardless of their sample size. If, however, censoring fractions are accounted for in the working variance matrix, as is done in the simplified estimator through the weight matrix h_{eff} , gains in precision follow.

The size of the gains in precision will vary according to the number of time intervals and the degree of autocorrelation in costs. In our simulations, more time intervals led to greater gains for the pooled improved and simplified estimators over the pooled working independence and stratified estimators, as did less correlation between time intervals. This is because the pooled improved and simplified estimators will have a greater relative efficiency when there is greater scope to share information across time intervals. Clearly, more time intervals allow for more sharing. Further, when there is less autocorrelation amongst intervals, there is more information in total and hence more information to share. The relative efficiency of the pooled improved to the pooled simplified estimator will decrease as the amount of potentially recoverable information lost to censoring decreases, *i.e.* as the censoring fraction decreases and as the width of the time intervals decreases. Thus, the efficiency of the pooled simplified estimator and the pooled improved estimator will be similar when there is a large number of narrow time intervals.

The simulation setting investigates the simplest case, where there is no difference between groups in terms of either cost or survival. This allows us to look at the simplest model, namely $E(C_{ik}) = \alpha_{0k} + \alpha_1 Z_{i2}$. If there is important mortality in the trial, then it is likely that as time progresses and fewer individuals survive, the between-treatment difference in cost will change over time, even if the difference in cost amongst survivors remains constant. One way of modelling this is to note that if $\Pr(\text{alive in interval } k \mid \text{group} = j) = p_{jk}$, we could set $E(C_{ik} \mid \text{group} = j) = p_{jk}(\alpha_{0k} + \alpha_1)$, *i.e.* we assume a constant between-group difference in costs amongst survivors. The difficulty with this is that the between-treatment difference in costs now depends on both α_{0k} and α_1 , and so the efficiency gained by sharing information between time intervals for α_1 may be offset by the fact that information is not shared for p or α_{0k} . However, if mortality does not differ by treatment group, then we can write $p_{jk} = p_k$, and the between-treatment difference in costs reduces to α_1 , and we can expect to regain some efficiency. An alternative approach when there is substantial mortality was illustrated in the example: it is not necessary to choose between stratifying and pooling; intermediate options are available too. In this example we allowed the between-treatment difference in costs to depend on year and quarter, but not their interaction (which in this case would result in a fully stratified model). Just as, if assuming a constant effect over time with longitudinal efficacy data were implausible, we would consider

modelling the shape of the effect over time (e.g. a linear trend) rather than immediately choosing to stratify, so with longitudinal cost data it will not always be necessary to stratify.

Our simplified estimator is not efficient. Rather, it is a compromise between the working-independence IPWGEE and the efficient estimator, and maintains some of the precision benefits whilst still being suitable for practical use. In terms of practical implementations of the simplified estimator, we have two cautions. Firstly, pooling across all time intervals at once may not be wise. This was observed in the CIDS study, where a more refined model including both seasonal and yearly variations was needed. Using an incorrect model may lead to inferences that are not only lacking in precision, but also biased. Chi-square tests for the appropriateness of pooling do exist, although they lack power. Rather than testing, it may be more instructive to plot the mean differences and their 95% confidence intervals, derived from the stratified model, and visually assess the nature of the variation across time. The second caution is that when there is a large number of time intervals, using an unstructured correlation may negate any precision benefits to pooling information across intervals, as this can result in estimating a large number of parameters. In GEEs, the exchangeable and autoregressive structures are popular choices. In our example, we used a more general option, letting the correlation between two observations depend on the lag (as a categorical covariate) and the time interval of the earlier observation. Code for estimating h_{eff} for exchangeable, lag-dependent, and interval and lag-dependent structures is available in the Appendix. Note, however, that whilst a poor working correlation structure will affect efficiency, it will not affect consistency.

Although the focus of this paper is on modelling, we have not considered parametric models for costs. Cost data typically have a very skewed distribution, and are difficult to model parametrically. In some cases, results can be sensitive to distributional assumptions (Nixon and Thompson, 2004). When parametric models are desired, log-transforms are a popular choice (see *e.g.* Liu *et al.*, 2007); however, if the aim of the analysis is to inform a cost-effectiveness evaluation, it is the between-treatment difference in mean costs that is of interest. This is not a parameter that is modelled when the distribution of costs is taken to be log-normal. Provided mean costs are not the focus of the analysis, however, parametric modelling has the potential to be more precise and more flexible than semi-parametric modelling. For example, shared random effects models allow costs and effects to be modelled jointly (Liu *et al.*, 2007, Gardiner *et al.*, 2006). Such joint modelling is a desirable feature, as a cost-effectiveness evaluation will always account for between-treatment differences in mean effectiveness, and will often need to trade off improved effectiveness vs. increased costs.

Another feature that, if accounted for, may improve precision, is the semi-continuous nature of cost data. Cost data typically have a number of values exactly equal to zero. See Liu (2009) and Tian and Huang (2007) for approaches to modelling this. Note, however, that since our proposed methods are semi-parametric, the semi-continuous nature of the data will not impact the consistency of the estimates.

In summary, we suggest that just as investigators have been opting for the partitioned BT estimator over the single-interval BT estimator when precision is a concern, so also they may consider non-stratified models combined with our simplified estimator when more precision than that offered by the partitioned BT estimator is required. It is reassuring that careful modelling appears to have a greater impact on precision than complex estimation procedures. Thus, when precision of estimation is a concern, we should aim to select a good model before worrying about efficient estimators.

Appendix: Code

```
tvarcoefg.Gnasimple.largeSS <-
function(Z,m,K,maxt,tint,X,delta,deltaC,p1,p2,structure){
# p1 is the number of covariates with time-varying coefficients
# p2 is the number of covariates with time-constant coefficients
# Z is the m*K*p matrix of covariates; the first p1 are assumed to
have coefficients that change over time, the last p2 have
coefficients that are constant over time
# m is the number of individuals
# tint is the vector defining the time intervals of interest
# X is the length of follow-up
# delta is the censoring indicator (delta=1 means no censoring)
# K is the number of time intervals
# deltaC is the m*K matrix of costs
# structure is the desired correlation structure for U and V, to
be chosen from "exchangeable", "lag", "lag+time"

  P <- p1 + p2
  Xstar <- as.numeric(X<maxt)*X + as.numeric(X>=maxt)*maxt
  deltabar <- 1-delta
  deltastar <- delta + as.numeric(X>=maxt)*deltabar
  c <- maxt+1
  if(p1>0){
    Z1 <- Z[,1:p1]
    Z1 <- array(Z1,dim=c(m,K,p1))
  }
  else Z1 <- array(dim=c(m,K,0))
  if(p2>0){
    Z2 <- Z[, (p1+1):p]
    Z2 <- array(Z2,dim=c(m,K,p2))
  }
  else Z2 <- array(dim=c(m,K,0))
  deltastark <- outer(X,tint,nge)
  deltastark <- sweep(deltastark,1,deltabar,"*")
  deltastark <- sweep(deltastark,1,delta,"+")
```

```
Xstark <- outer(X,tint,arraymin)

km <- survfit(Surv(X,deltabar)~1)
G <- km$surv[km$n.event>0]
time.g <- basehaz(coxph(Surv(X,deltabar)~1))$time
nt.g <- length(time.g)
ph.g <- coxph(Surv(X,deltabar)~1)
H0.g <- basehaz(ph.g,centered=FALSE)$hazard
h0.g <- H0.g - c(0,H0.g[1:(nt.g-1)])
haz.g <- matrix(rep(h0.g,m),nrow=m,byrow=TRUE)
Gminus <- c(1,G[1:(nt.g-1)])

deltastark <- deltastark[,2:(K+1)]
Xstark <- Xstark[,2:(K+1)]
indexk <- apply(Xstark,1:2,indexfne,time.g)

index.Gint <- rep(0,nt.g)
  for(i in 1:nt.g){
    index.Gint[i] <- max(1,sum(as.numeric(tint <= time.g[i])))
  }

Gnak <- array(dim=c(m,K))
for(k in 1:K){
  for(i in 1:m){
    if(indexk[i,k]>0) Gnak[i,k] <- G[indexk[i,k]]
    else Gnak[i,k] <- 1
    if(Gnak[i,k]==0 & deltastark[i,k]==0) Gnak[i,k] <- 1
  }
}

heff <- array(dim=c(m,K,K))
for(i in 1:m) heff[i,,] <- diag(rep(1,K))
  soln <-
heffGEEsolve(deltaC,deltastark,Gnak,Z1,Z2,p1,p2,K,m,heff)

alphav.1 <- soln$alphav; alphaf.1 <- soln$alphaf

# More efficient estimators

# alphak different
margin.mean <- array(dim=c(m,K))
for(k in 1:K){
  margin.mean[,k] <-
array(Z1[,k,],dim=c(m,p1))%*%array(alphav.1[k,],dim=p1) +
array(Z2[,k,],dim=c(m,p2))%*%array(alphaf.1, dim=p2)
}

Ustore1 <-
Uest.condalive(m,K,p,deltastar,deltastark,Xstar,deltaC,Z[,1,2],margin.mean,Gnak,tint,structure)
U <- Ustore1$U

expected <- sweep(Ustore1$cond.mean,2:3,margin.mean,"-")
  remove(Ustore1)
```

Pullenayegum and Willan: Models for Censored Longitudinal Cost Data

```

heff <- array(dim=c(m,K,K))
for(i in 1:m){
  heff[i,,] <- solve(U[Z[i,1,2]+1,1,,] +
apply(sweep(U[Z[i,1,2]+1,index.Gint,,],1,(haz.g[i,]/Gminus),"*"),2
:3,sum),tol=10^(-100))
}

remove(U)
heff <- heff/sum(heff[1,,])

soln <-
heffGEEsolve(deltaC,deltastark,Gnak,Z1,Z2,p1,p2,K,m,heff)

return(list(alphaf.1=alphaf.1,alphaf.1=alphaf.1,alphaf.eff=soln
$alphaf,alphaf.eff=soln$alphaf))
}

correst <- function(y,subj,weight,nsubj,K,Zbin,structure){
# assumes balanced
Vy <- array(dim=c(2,K))
Vy[1,] <- apply(matrix(y*y*weight,ncol=K)[Zbin==1,],2,mean)
Vy[2,] <- apply(matrix(y*y*weight,ncol=K)[Zbin==2,],2,mean)

if(K==2) structure <- "lag"

for(i in 1:nsubj){
  yi <- y[subj==i]
  nobs <- length(yi)
  yi <- yi/sqrt(Vy[Zbin[i],1:nobs])
  xprod <- yi%*%t(yi)
  tri <- array(0,dim=c(nobs,nobs))
  for(j in 1:nobs) for(k in 1:nobs) if(j >=k) tri[j,k] <- 1
  tri <- as.vector(tri)
  wgt <- rep(weight[subj==i],nobs)
  xprod <- as.vector(xprod)
  xprod <- xprod[tri==1]
  wgt <- wgt[tri==1]
  t1 <- rep(1:nobs,nobs); t1 <- t1[order(t1)]; t1 <- t1[tri==1]
  t2 <- rep(1:nobs,nobs); t2 <- t2[tri==1]
  lag <- array(dim=c(nobs,nobs))
  for(j in 1:nobs) for(k in 1:nobs) lag[j,k] <- abs(j-k)
  lag <- as.vector(lag); lag <- lag[tri==1]
  if(i==1){ w <- xprod; time1 <- t1; time2 <- t2; weight.w <-
wgt; id <- rep(i,length(wgt)); lagw <- lag}
  else{ w <- c(w,xprod); time1 <- c(time1,t1); time2 <-
c(time2,t2); weight.w <- c(weight.w,wgt); id <-
c(id,rep(i,length(wgt))); lagw <- c(lagw,lag)}
}

# Variances - down the diagonal
v <- rep(0,K)

```

```
for(k in 1:K) v[k] <- mean((w*weight.w)[lagw==0 & time1==k])

response <- w*weight.w
x <- sqrt(v[time1]*v[time2])

if(structure=="exchangeable"){
  m <- lm(I(w/x) ~ 1, subset=lagw>0, weight=weight.w)
  print(summary(m))
  alpha <- m$coefficients[1]

  V <- array(alpha,dim=c(2,K,K))
  V[1,,] <- V[1,,] + diag(rep(1-alpha,K))
  V[1,,] <-
diag(sqrt(Vy[1,]))%*%diag(sqrt(v))%*%V[1,,]%*%diag(sqrt(v))%*%diag
(sqrt(Vy[1,]))
  V[2,,] <- V[2,,] + diag(rep(1-alpha,K))
  V[2,,] <-
diag(sqrt(Vy[2,]))%*%diag(sqrt(v))%*%V[2,,]%*%diag(sqrt(v))%*%diag
(sqrt(Vy[2,]))
}

if(structure=="lag"){
  lag <- array(dim=max(lagw))
  for(l in 1:max(lagw)){
    lag[l] <- mean(response[lagw==l])
  }
  C <- array(dim=c(K,K))
  for(j in 1:K){
    for(k in 1:K){
      if(j==k) C[j,k] <- 1
      else C[j,k] <- lag[abs(j-k)]
    }
  }
  V <- array(dim=c(2,K,K))
  V[1,,] <- diag(sqrt(Vy[1,]))%*%C%*%diag(sqrt(Vy[1,]))
  V[2,,] <- diag(sqrt(Vy[2,]))%*%C%*%diag(sqrt(Vy[2,]))
}

if(structure=="lag+time"){
  tmax <- time1*as.numeric(time1>time2) +
time2*as.numeric(time2>=time1)
  model <- lm(response ~ factor(lagw) + factor(tmax),
subset=lagw>0)

  print(summary(model))

  C <- array(dim=c(K,K))
  for(j in 1:K){
    for(k in 1:K){
      newdata <- cbind(abs(j-k),max(j,k))
      newdata <- as.data.frame(newdata); names(newdata) <-
c("lagw","tmax")
      if(j==k) C[j,k] <- 1
      else C[j,k] <- predict(model,newdata=newdata)
    }
  }
}
```


Pullenayegum and Willan: Models for Censored Longitudinal Cost Data

```

V <- array(dim=c(2,K,K))
V[1,,] <- diag(sqrt(Vy[1,]))%*%C%*%diag(sqrt(Vy[1,]))
V[2,,] <- diag(sqrt(Vy[2,]))%*%C%*%diag(sqrt(Vy[2,]))
}

return(V)
}

Uest.condalive <-
function(m,K,p,deltastar,deltastark,Xstar,deltaC,Z,margin.mean,Gna
k,tint,structure){

# Z here is assumed to be a single covariate
# Estimates  $U = E(\text{var}(\text{zeta}|\text{history to time } u)|Z)$ 

deltaC <- data.matrix(deltaC)

# Calculate  $V=\text{var}(\text{deltaC}|Z)$  by smoothing
V <- array(0,dim=c(2,K,K))
nlevelsz <- dim(table(Z))
if(nlevelsz>2) Zbin <-
as.numeric(cut(Z,breaks=quantile(Z,probs=c(0,0.5,1)),include.lowes
t=TRUE))
else Zbin <- 1 + round((Z-min(Z))/(max(Z)-min(Z)))

deltaC.std <- deltaC - margin.mean
weight <- deltastark/Gnak

V <- array(0,dim=c(2,K,K))
V[,2:K,2:K] <- correst(as.vector(deltaC.std[,2:K]),rep(1:m,K-
1),as.vector(weight[,2:K]),m,K-1,Zbin,structure)
V[,1,1] <-
tapply(deltaC.std[,1]*deltaC.std[,1]*weight[,1],Zbin,mean)

# Calculate conditional expectations

cond.mean <- array(dim=c(m,K+1,K))
cond.mean[,1,] <- margin.mean

for(k1 in 1:K){
  for(k2 in 1:k1){
    cond.mean[,k1+1,k2] <- deltaC[,k2]
  }
}

for(k1 in 1:(K-1)){
  for(k2 in (k1+1):K){
    if(sum(deltastark[,k2])>10){
      obs <- deltaC[,k2]
      weight <- deltastark[,k2]*Gnak[,k1]/Gnak[,k2]
      if(k1>1) runmean <- apply(deltaC[,1:k1],1,mean)
      else runmean <- deltaC[,k1]
      runmean.cat <- rep(0,m)
    }
  }
}

```

```

runmean.cat[runmean==0] <- 0
runmean.cat[runmean>0] <-
as.numeric(factor(cut(runmean[runmean>0],breaks=quantile(runmean[r
unmean>0],probs=((0:5)/5)),include.lowest=TRUE)))
if(sum(weight>0)>40) model.cmean <- lm(obs~runmean.cat
+ Zbin,subset=(weight>0& Xstar>tint[k1+1]),na.action=na.omit,
weight=weight)
else model.cmean <- lm(obs~ runmean.cat +
Zbin,subset=(weight>0 & Xstar>tint[k1+1]),na.action=na.omit,
weight=weight)
predict.data <-
data.frame(cbind(obs,Xstar,runmean,runmean.cat,Zbin,weight))
names(predict.data) <-
c("obs","Xstar","runmean","runmean.cat","Zbin","weight")

cond.mean[Xstar>=tint[k1+1],k1+1,k2] <-
predict(model.cmean,newdata=predict.data[Xstar>=tint[k1+1],])
cond.mean[Xstar <= tint[k1+1],k1+1,k2] <- 0
}
else cond.mean[,k1+1,k2] <- margin.mean[,k2]
cond.mean[,k1+1,k2] <- cond.mean[,k1+1,k2]*
as.numeric(Xstar >= tint[k1+1])
}
}

U <- array(0,dim=c(2,K+1,K,K))

# Conditional covariance
for(k1 in 1:(K-1)){
  alive <- as.numeric(Xstar>tint[k1+1])
  deltaC.std <- deltaC - cond.mean[,k1+1,]
  deltaC.std <- deltaC.std[alive==1,(k1+1):K]
  weight <-
sweep(deltastark/Gnak,1,Gnak[,k1],"*")[alive==1,(k1+1):K]
  nsubj <- sum(alive)
  if(k1<(K-1)) c <-
correst(as.vector(deltaC.std),rep(1:nsubj,K-
k1),as.vector(weight),nsubj,K-k1,Zbin[alive==1],structure)
  else c <-
tapply(deltaC.std*deltaC.std*weight,Zbin[alive==1],mean)

  U[,k1+1,(k1+1):K,(k1+1):K] <- mean(alive)*c
}

U[,1,,] <- V
return(list(U=U,V=V,cond.mean=cond.mean))
}

heffGEESolve <-
function(deltaC,deltastark,Gnak,Z1,Z2,p1,p2,K,m,heff){
  ZZff <- array(0,dim=c(p2,p2))
  ZZfv <- array(0,dim=c(K,p2,p1))
  ZZvf <- array(0,dim=c(K,p1,p2))
  ZZvv <- array(0,dim=c(K,K,p1,p1))
  for(k in 1:K){
    for(i in 1:m){

```

```

        for(k2 in 1:K){
            ZZvv[k,k2,,] <- ZZvv[k,k2,,] +
            Z1[i,k,]*(deltastark[i,k2]*heff[i,k,k2]/Gnak[i,k2])*Z1[i,k2,]
        }
        ZZfv[k,,] <- ZZfv[k,,] +
        drop((t(Z2[i,,])%*%(deltastark[i,k]*heff[i,,k]/Gnak[i,k])))%*%t(Z1[i,k,]))
        ZZvf[k,,] <- ZZvf[k,,] +
        Z1[i,k,]%*%t(drop(t(deltastark[i,]*heff[i,k,]/Gnak[i,])%*%Z2[i,,]))
    )
}
for(i in 1:m){
    ZZff <- ZZff +
    t(array(sweep(array(Z2[i,,],dim=c(K,p2)),1,(deltastark[i,]/Gnak[i,]),"***"),dim=c(K,p2))%*%heff[i,,]%*%Z2[i,,])
}
Xfv <- array(0,dim=c(p2,p1*K)); Xvv <- array(0,dim=c(p1*K,p1*K));
Xff <- ZZff; Cv <- array(dim=c(K,p1)); Cf <- array(0,dim=p2)
if(p2>0){for(i in 1:p2){Xfv[i,] <- as.vector(ZZfv[,i,])}}
Xvf <- array(ZZvf[1,,],dim=c(p1,p2))
for(k in 1:K){
    if(k>1) Xvf <- rbind(Xvf,array(ZZvf[k,,],dim=c(p1,p2)))
    for(k2 in 1:K){
        Xvv[((k-1)*p1+1):(k*p1),((k2-1)*p1+1):(k2*p1)] <-
        ZZvv[k,k2,,]
    }
    if(p1>1) Cv[k,] <-
    drop(array(array(sweep(Z1[,k,],1,apply(heff[, ,k]*(deltaC*deltastark/Gnak),1,sum),"***"),dim=c(m,p1)),1,sum))
    else Cv[k,] <-
    sum(array(sweep(array(Z1[,k,],dim=c(m,p1)),1,apply(heff[, ,k]*(deltaC*deltastark/Gnak),1,sum),"***"),dim=c(m,p1)))
}
apply(sweep(Z2,1:2,apply(sweep(heff,c(1,3),(deltastark*deltaC/Gnak),"***"),c(1,2),sum),"***"),3,sum)
deltaC <- data.matrix(deltaC)
for(i in 1:m){
    Cf <- Cf +
    drop(t(Z2[i,,])%*%(heff[i,,]%*%(deltastark[i,]*deltaC[i,]/Gnak[i,]))))
}
Xmat <- cbind(rbind(Xvv,Xfv),rbind(Xvf,Xff))
Y <- c(as.vector(t(Cv)),Cf)
valpha <- solve(Xmat,Y,tol=10^(-100))
if(p1>0) alphav <- matrix(valpha[1:(K*p1)],nrow=K,byrow=TRUE)
else alphav <- array(dim=c(K,0))
if(p2>0) alphaf <- valpha[(K*p1+1):(K*p1+p2)]
else alphaf <- array(dim=0)
return(list(alphaf=alphaf,alphav=alphav))
}

indexfne <- function(x,y){return(sum(as.numeric(y<=x)))}
nge <- function(a,b){ return(as.numeric(a>=b))}
arraymin <- function(x,y){return(x*(x<y) + y*(y<=x))}

```

References

- Bang H and Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 2000; **87**: 329-343.
- Baser O, Gardiner JC, Bradley CJ, Yuce H, Given C. Longitudinal analysis of censored medical cost data. *Health Economics* 2006; **15**: 513-525.
- Connolly SJ, Gent M, Roberts RS, Dorian P, Roy D, Sheldon RS, Mitchell LB, Green MS, Klein GJ, O'Brien B. Canadian implantable defibrillator study (CIDS): a randomized trial of the implantable cardioverter defibrillator against amiodarone. *Circulation* 2000; **101**: 1297-1302.
- Gardiner JC, Luo Z, Bradley CJ, Sirbu CM, Given CW. A dynamic model for estimating changes in health status and costs. *Statistics in Medicine* 2006; **25**: 3648-3667.
- Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53**:419-34.
- Lin DY. Linear regression analysis of censored medical costs. *Biostatistics* 2000; **1**: 35-47.
- Lin DY. Regression analysis of incomplete medical cost data. *Statistics in Medicine* 2003; **22**: 1181-1200.
- Liu L, Wolfe RA, Kalbfleisch JD. A shared random effects model for censored medical costs and mortality. *Statistics in Medicine* 2007; **26**: 136-155.
- Liu L. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* 2009; **28**:972-986.
- Nixon RM, Thompson SG. Parametric modelling of cost data in medical studies. *Statistics in Medicine* 2004; **23**:1311-1331.
- O'Brien BJ, Connolly SJ, Goree R, Blackhouse G, Willan AR, Yee R, Roberts RS, Gent M. Cost-effectiveness of the implantable cardioverter defibrillator: results from the Canadian Implantable Defibrillator Study (CIDS). *Circulation* 2000; **103**: 1416-1421.
- O'Hagan A, Stevens JW. On estimators of medical costs with censored data. *Health Economics* 2004; **23**: 615-625.
- Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**: 1033-1048.
- Pullenayegum EM, Willan AR. Semi-parametric models for cost-effectiveness analysis: improving the efficiency of estimation from censored data. *Statistics in Medicine* 2007; **26**: 3274-3299.
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**: 106-121.

- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**: 846-866.
- Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell NP, Dietz K, Farewell VT, *AIDS Epidemiology: Methodological Issues*. Birkhäuser: Boston, 1992; pp. 297-331.
- Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regressions models with missing data. *Journal of the American Statistical Association* 1995; **90**: 122-129.
- Tian L, Huang J. A two-part model for censored medical cost data. *Statistics in Medicine* 2007; **26**: 4273-4292.
- Zhao H, Bang H, Wang H, Pfeifer PE. On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine* doi 10.1002/sim.2882.