Published in final edited form as:

Biometrics. 2011 March ; 67(1): 1–7. doi:10.1111/j.1541-0420.2010.01416.x.

Marginal models for clustered time to event data with competing risks using pseudovalues

Brent R. Logan^{*}, Mei-Jie Zhang, and John P. Klein

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, U.S.A

Summary

Many time-to-event studies are complicated by the presence of competing risks and by nesting of individuals within a cluster, such as patients in the same center in a multi-center study. Several methods have been proposed for modeling the cumulative incidence function with independent observations. However, when subjects are clustered, one needs to account for the presence of a cluster effect either through frailty modeling of the hazard or subdistribution hazard, or by adjusting for the within-cluster correlation in a marginal model. We propose a method for modeling the marginal cumulative incidence function directly. We compute leave-one-out pseudo-observations from the cumulative incidence function at several time points. These are used in a generalized estimating equation to model the marginal cumulative incidence curve, and obtain consistent estimates of the model parameters. A sandwich variance estimator is derived to adjust for the within cluster correlation. The method is easy to implement using standard software once the pseudo-values are obtained, and is a generalization of several existing models. Simulation studies show that the method works well to adjust the standard error for the within cluster correlation. We illustrate the method on a dataset looking at outcomes after bone marrow transplantation.

Keywords

Cumulative incidence; Marginal model; Pseudo-observations; Generalized estimating equations; Sandwich variance; Clustered data

1. Introduction

Many time-to-event studies in biomedical applications are complicated by the presence of competing risks, in which failure of one type precludes observation of failure of another type. The standard method for modeling this type of data is to apply Cox regression models to each cause-specific hazard function. More recently, several researchers have proposed direct regression models for the cumulative incidence function, defined as the probability of failing from cause 1 before time t (Fine and Gray, 1999; Klein and Andersen, 2005; Scheike, Zhang, and Gerds, 2008). These methods have been developed assuming independence between subjects.

In some clinical study designs the subjects may be clustered or correlated rather than independent. For example, in multi-center studies patients within the same center may be

^{*}blogan@mcw.edu.

Supplementary Materials Web appendix A referenced in Section 4 is available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

correlated due to a latent center effect. Alternatively, in matched cohort studies patients within the same matched stratum are correlated. In these study designs, one needs to account for the correlation induced by the clustered structure of the data.

This research is motivated by an example in hematopoietic cell transplantation, in which researchers are interested in modeling the cumulative incidence of transplant-related mortality (death in the absence of relapse of disease) as a function of patient and transplant characteristics. These transplants are occurring at a number of different transplant centers, and often the transplant characteristics, such as the choice of graft type, conditioning regimen, or graft-versus-host disease prophylaxis, are associated with center. Other unmeasured variables such as supportive care, could also contribute to a center effect. An analysis of the effect of patient and transplant characteristics should account for the clustering of patients within a transplant center.

Chen et al. (2007) propose a method for adjusting the variance of a cumulative incidence estimator when the observations are clustered, as well as two-sample tests which are analogs of Gray's test and Pepe's test. Katsahian et al. (2008) propose a frailty model for the subdistribution hazard in a regression framework. We propose instead a regression model for the marginal cumulative incidence function. We use the pseudo-observation framework of Klein and Andersen (2005), which relies on generalized estimating equations to estimate the parameters of the marginal model. The variance is corrected for clustering using a sandwich estimate. This method is easy to implement using existing software once the pseudovalues have been obtained.

The paper is organized as follows. In Section 2 we lay out the notation and describe the proposed method and its relationship to other existing procedures. In Section 3 we show the results of a simulation study to investigate the properties of the proposed procedure, in terms of bias and SE. In Section 4 we illustrate the procedure with the example in hematopoietic cell transplantation described previously. Finally, in Section 5 we summarize our conclusions.

2. Methods

2.1 Notation

The data consist of m clusters with n_i subjects in cluster i , for a total of n observations. The observation for subject j in cluster i is $(X_{ij}, \delta_{ij}, \varepsilon_{ij}, Z_{ij})$. Here X_{ij} is the on-study time given by $X_{ij} = (T_{ij} \wedge C_{ij})$, where T_{ij} is the event time and C_{ij} is the censoring time. Also, $\delta_{ij} = I(T_{ij} \leq C_{ij})$ is the event indicator and $\varepsilon_{ij} \in \{1, \dots, k\}$ denotes the failure type, observable if an event has occurred prior to C_{ij} . Finally, Z_{ij} refers to the covariate vector. We assume that the C_{ij} 's are a random sample from a distribution with survival function $G(t)$, and that C_{ij} is independent of T_{ij} .

Without loss of generality we will consider models for the marginal cumulative incidence function for cause 1, $F_1(t) = P(T \leq t, \varepsilon = 1)$. Define $N_{ij}(t) = I(X_{ij} \leq t) I(\varepsilon_{ij} = 1) \delta_{ij}$, and let $N(t) = \sum_{i,j} N_{ij}(t)$. Here $N(t)$ is the number of type 1 failures observed up to time t . Denote the number at risk at time t by $R(t)$.

The cumulative incidence estimate is defined as $\hat{F}_1(t) = \int_0^t \hat{S}(u^-) R^{-1}(u) dN(u)$, where $\hat{S}(u^-)$ is the Kaplan-Meier estimate of event-free survival, in which the patient has not experienced either cause 1 or the competing event. Chen et al. (2007) show that $\hat{F}_1(t)$ is a consistent estimate of $F_1(t)$ even when the data are clustered. Also, this estimate can alternatively be written using an inverse probability of censoring weighting (IPCW) formulation (Scheike et al., 2008) as

$$\widehat{F}_1(t) = \frac{1}{n} \sum_{i,j} \frac{N_{ij}(t)}{\widehat{G}(X_{ij})}, \quad (1)$$

where $\widehat{G}(t)$ is the Kaplan-Meier estimate of the censoring survival distribution obtained by treating censored observations as events. We are interested in modeling the relationship between the marginal cumulative incidence function and patient-specific covariates $\mathbf{Z}_{ij}$. To do this, we will use the pseudo-value approach of Andersen, Klein, and Rosthøj (2003) and Klein and Andersen (2005) as described in the next subsection.

2.2 Pseudo-value approach

The pseudo-value at time t is defined by $Y_{ij}(t) = n\widehat{F}_1(t) - (n-1)\widehat{F}_1^{-(ij)}(t)$, where $\widehat{F}_1^{-(ij)}(t)$ is the cumulative incidence estimate obtained by omitting the j th patient in the i th cluster. In Appendix A we show using the IPCW formulation for $\widehat{F}_1(t)$ that the pseudo-value can be expressed as

$$Y_{ij}(t) = \frac{N_{ij}(t)}{G(X_{ij})} + \int_0^{X_{ij}} \frac{P(T \leq t, \varepsilon=1 | T \geq u)}{G(u)} dM_{ij}^c(u) + O_p(n^{-1/2}), \quad (2)$$

where $M_{ij}^c(t) = I(X_{ij} \leq t)(1 - \delta_{ij}) - \int_0^t I(X_{ij} \geq u) \lambda^c(u) du$, is the martingale corresponding to the censoring process for observation (i, j) , with censoring hazard function $\lambda^c(u) = -d \log\{G(u)\} / du$. We point out that a similar derivation in the independent observation case has been formalized using influence functions by Graw et al. (2009).

Using this formulation, we point out several properties of the pseudo-values.

- (C1) $Y_{ij}(t)$ is approximately independent of $Y_{kl}(t)$ for clusters $i \neq k$ as $n \rightarrow \infty$. This follows because $Y_{ij}(t)$ only depends on X_{ij} and $dM_{ij}^c(t)$ as n approaches ∞ .
- (C2) The expected value of the pseudo-value given the covariate value $\mathbf{Z}_{ij}$ converges to

$$\lim_{n \rightarrow \infty} E\{Y_{ij}(t) | \mathbf{Z}_{ij}\} = F_1(t | \mathbf{Z}_{ij}).$$

This follows since the second term in (2) is a martingale with mean 0 independent of $\mathbf{Z}_{ij}$, and because $E\{N_{ij}(t)/G(X_{ij})\} = F_1(t | \mathbf{Z}_{ij})$. (Scheike et al., 2008).

Suppose we are interested in modeling the marginal cumulative incidence function as a function of covariates, according to the model $g\{F_1(t | \mathbf{Z})\} = \boldsymbol{\beta}' \mathbf{Z}$. Alternatively, we can write $\mu = F_1(t | \mathbf{Z}) = g^{-1}(\boldsymbol{\beta}' \mathbf{Z})$. As is described in Andersen, Klein, and Rosthøj (2003) and Klein and Andersen (2005), the pseudo-values can be used in a generalized estimating equation (GEE) framework to estimate $\boldsymbol{\beta}$. Define the vector $\mathbf{Y}_i = \{Y_{ij}(t_r), \text{ for } j = 1, \dots, n_i; r = 1, \dots, q\}$ for a set of time points $(t_1, \dots, t_q)$, where $\mathbf{Y}_i$ has dimension $q \cdot n_i$. Similarly define the vector $\boldsymbol{\mu}_i = \{\mu_{ij}(t_r), \text{ for } j = 1, \dots, n_i; r = 1, \dots, q\}$ where $\mu_{ij}(t_r) = F_1(t_r | \mathbf{Z}_{ij})$. The GEE is given by

$$\sum_i \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \sum_i \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0},$$

where V_i is a $q \cdot n_i \times q \cdot n_i$ working correlation matrix. Because of the properties (C1) and (C2), $U_i(\beta)$ are asymptotically independent with mean $\mathbf{0}$ given covariates. Therefore, using the results of Liang and Zeger (1986), under regularity conditions, $\sqrt{m}(\widehat{\beta} - \beta)$ converges to a multivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix Σ . A consistent estimate of Σ even with a misspecified working correlation matrix is given by the sandwich variance

$$\widehat{\Sigma} = \widehat{I(\beta)}^{-1} \widehat{\text{Var}\{U(\beta)\}} \widehat{I(\beta)}^{-1}, \quad (3)$$

with $I(\beta) = \sum_i (\partial \mu_i / \partial \beta)' V_i^{-1} (\partial \mu_i / \partial \beta)$, and

$$\widehat{\text{Var}\{U(\beta)\}} = \sum_i U_i(\beta) U_i(\beta)'. \quad (4)$$

If we use a working correlation matrix V_i which is block diagonal with blocks denoted V_{ij} , then the GEE simplifies to that of the non-clustered data situation,

$$\sum_{i,j} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' V_{ij}^{-1} (Y_{ij} - \mu_{ij}) = \sum_{i,j} U_{ij}(\beta) = 0. \quad (5)$$

Note with the GEE in (5) obtained ignoring clustering, the model-based information matrix

$I(\beta)$ simplifies to $I(\beta) = \sum_{i,j} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' V_{ij}^{-1} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)$. However, the empirical variance estimate based on the GEE in (5), given by

$$\widehat{\text{Var}\{U(\beta)\}} = \sum_{ij} U_{ij}(\beta) U_{ij}(\beta)', \quad (6)$$

does not yield a consistent estimate of $\text{Var}\{U(\beta)\}$ when the observations are correlated within clusters. Therefore, it is important in the presence of correlation within cluster that the middle part of the sandwich variance estimate reflect the clustered data structure, as in (4). This is illustrated further in the simulation studies.

In general, it may be difficult to specify an appropriate working correlation matrix, because of the complexity of the correlation structure. We have two sources of correlation here. Each individual's pseudovalue is computed at multiple time points. Furthermore, we assume that there is within cluster correlation between patients in the same cluster. We can still obtain consistent estimates of β using an independence working correlation matrix, $V_i = \text{diag}\{1\}$. In

this case, U_i simplifies to $\sum_{j=1}^{n_i} \sum_{r=1}^q (\partial \mu_{ijr} / \partial \beta) \{Y_{ij}(t_r) - \mu_{ijr}\}$ and $I(\beta)$ simplifies to $\sum_i \sum_{j=1}^{n_i} \sum_{r=1}^q (\partial \mu_{ijr} / \partial \beta)' (\partial \mu_{ijr} / \partial \beta)$. Note that these marginal models can be fit using standard software once the pseudovalue are obtained; see Klein et al. (2008) for programs to compute the pseudovalue.

2.3 Special Cases

This pseudo-value regression model was proposed for the general case of modeling the marginal cumulative incidence curve $g\{F_1(t|\mathbf{Z})\} = \beta' \mathbf{Z}$ as a function of covariates. Note that when there are no clusters so that all the patients are independent, the general framework simplifies to that of Klein and Andersen (2005). Furthermore, as they point out, if the complementary log-log (CLL) transformation model is used, this is analogous to a proportional hazards model on the subdistribution hazard function as used in Fine and Gray (1999).

For the clustered data situation, the model proposed can be shown to be a generalization of several other models as well. When there is no competing outcome, the model can be easily adapted to use the pseudo-value defined for the survival probability instead of the cumulative incidence probability. For example, write the Kaplan-Meier estimate using an IPCW framework (Satten and Datta, 2001) as $1 - \hat{S}(t) = n^{-1} \sum_{i,j} \hat{G}^{-1}(X_{ij}) \tilde{N}_{ij}(t)$, where $\tilde{N}_{ij} = I(X_{ij} \leq t) \delta_{ij}$. Then the pseudo-value can be shown to be equal to

$$Y_{ij}(t) = \frac{\tilde{N}_{ij}(t)}{G(X_{ij})} + \int_0^{X_{ij}} \frac{P(T \leq t, \delta=1 | T \geq u)}{G(u)} dM_{ij}^c(u) + O_p(n^{-1/2}),$$

so that properties analogous to (C1) and (C2) hold. If we then use a complementary log-log transformation in the GEE framework, the resulting model is similar to the marginal Cox model of Lee, Wei, and Amato (1992) and Spiekerman and Lin (1998).

Sometimes researchers may be interested in the cumulative incidence probability at a single time point t . If there is no censoring prior to that time point, the pseudo-value is simply a binary indicator of whether the event of interest has occurred prior to t . In this case, the marginal regression model using a logit link function is simply the clustered binary regression model of Liang and Zeger (1986). Hence the pseudo-value regression model with a single time point can be seen as a generalization of the logistic regression model for correlated binary data, in the presence of censoring. Note that in the non-clustered case, this relationship between the pseudo-value regression method and logistic regression has been discussed in Klein et al. (2007).

3. Simulation Study

In order to conduct a simulation study to examine the properties of our estimators, we need to be able to generate clustered competing risks data which follow a marginal model. In particular we are interested in generating data which follow a marginal proportional subdistribution hazards model, which can be estimated using pseudo-values with a complementary log-log transformation. This can be done by noting the relationship between marginal models and frailty models for the subdistribution hazard function. Katsahian et al. (2006) assume a proportional hazards representation where the subdistribution hazard given the frailty ω is defined as $\eta_1(t|\mathbf{Z}, \omega) = \eta_0(t)\omega \exp(\beta'\mathbf{Z})$, for $\omega > 0$. It is important to note that a proportional hazards model for the subdistribution hazard conditional on the frailty does not necessarily translate into a proportional hazards model for the marginal subdistribution hazard. This is also the case for the non competing risk setting, in which a specific frailty distribution (positive stable frailty) is necessary in order to obtain a marginal proportional hazards model. Here we show that the use of a positive stable frailty also results in a marginal model which has proportional subdistribution hazards.

Under the frailty model for the subdistribution hazard, the cumulative incidence function has the form

$$F_1(t|\mathbf{Z}, \omega) = 1 - \{1 - F_{01}(t)\}^{\omega \exp(\beta' \mathbf{Z})}.$$

The marginal cumulative incidence function can be written in terms of the Laplace transform $L(s) = E\{\exp(-sW)\}$ as $F_1(t|\mathbf{Z}) = 1 - L[-\ln\{1 - F_{01}(t)\} \exp(\beta' \mathbf{Z})]$.

Since the Laplace transformation of the standard positive stable distribution with parameter α is $L(s) = \exp(-s^\alpha)$, then $F_1(t|\mathbf{Z}) = 1 - \exp[-\{-\ln(1 - F_{01}(t))\}^\alpha \exp(\alpha\beta' \mathbf{Z})]$, and $\ln[-\ln\{1 - F_1(t|\mathbf{Z})\}] = \alpha \ln[-\ln\{1 - F_{01}(t)\}] + \alpha\beta' \mathbf{Z}$, resulting in a marginal proportional hazards representation with parameter vector rescaled according to $\tilde{\beta} = \alpha\beta$. This is exactly the same result as occurs in the non-competing risk setting (Hougaard, 2000).

Using this result, we can design the simulation study as follows. We consider number of clusters to be 50, 150, or 300, and number of observations per cluster n_i equal to 2, 4, or 8. For each cluster a random effect ω is generated from a positive stable frailty distribution with parameter α equal to 1.0 (independence within clusters), 0.5, or 0.25. These values of α would correspond to values of Kendall's τ equal to 0, 0.5, or 0.75 in the survival setting. Conditional on the frailty, competing risk data are generated according to the models

$$F_1(t|\omega, \mathbf{Z}) = 1 - [1 - p\{1 - \exp(-t)\}]^{\omega \exp(\beta' \mathbf{Z})}$$

$$F_2(t|\omega, \mathbf{Z}) = (1 - p)^{\omega \exp(\beta' \mathbf{Z})} [1 - \exp\{-t \exp(\beta' \mathbf{Z})\}]$$

This results in a marginal model for cause 1 which has proportional subdistribution hazards as described in the previous section. Given the cluster frailty ω and covariate $\mathbf{Z}$, data were randomly generated by selecting the failure cause randomly with probability $F_1(\infty|\omega, \mathbf{Z})$ for cause 1, then generating from the conditional distribution of the time to event given the failure cause and ω using an inverse CDF technique.

Values of p are selected to result in 25, 50, or 75% events of type 1. Independent censoring times are generated from an exponential distribution resulting in censoring percentages of 0%, 25%, or 50%. Three binary covariates are generated for each model, and each is generated differently to illustrate the impact of clustering on different types of covariates. The first covariate Z_1 is generated so that each cluster has equal frequencies of both levels of the covariate. This imitates a matched pairs design when $n_i = 2$ for each cluster, in which each cluster has both levels of the covariate represented. The second covariate Z_2 is generated independently from the cluster with probability 0.5 for each level. The third covariate Z_3 is a cluster level covariate in which the covariate value for the cluster is generated randomly with probability 0.5 for each level. Here all individuals within a cluster have the same value of covariate Z_3 . The parameters associated with these are $\beta = (0.5, -0.2, -0.3)/\alpha$, so that the parameters of the marginal model are the same across all simulations. For each dataset generated, we estimated the parameters of the marginal model using an independent working covariance matrix with 5 time points spaced approximately equally across the event times. A total of 5,000 Monte Carlo replicates were used, and the simulations were implemented using a program written in C.

First we examine the performance of the procedure in terms of the bias of the parameter estimates, since the estimating equations are only asymptotically unbiased. Boxplots of the relative bias ($= \text{Bias}/\beta$) are shown in Figure 1 for a partitioning of the simulation

configurations, according to number of clusters m , observations per cluster n_i , % type I events, and censoring %. The boxplots represent the range of values across all remaining simulation settings, i.e. across α and the covariate number. Note that although this is not shown in the figure, the relative bias did not vary with the frailty parameter α .

These results confirm that the relative bias gets smaller as the number of clusters or the number of observations per cluster gets larger. This bias is generally a small fraction of the parameter value, less than 5-10%, except for when $m = 50$, $n_i = 2$, 25% type I events, and 50% censoring. These problems are likely caused by a very small expected frequency of type I events. In general, the bias shouldn't be a concern as long as there is a reasonable frequency of type I events.

Next we consider the variance of the parameter estimates. Three estimators are compared: ignoring the clustering using the empirical variance formula in (6), accounting for the clustering using the empirical variance formula in (4), or an alternative bootstrap variance estimate. This bootstrap variance estimate was obtained by resampling clusters, so that every individual within a selected cluster was included. For each bootstrap sample, the regression model was refit and the parameter estimate $\hat{\beta}^*$ was obtained. The bootstrap variance estimate was the sample variance of the $\hat{\beta}^*$'s across 100 bootstrap samples. The actual standard deviation (SD) of the $\hat{\beta}$'s was estimated using the Monte Carlo replicates. For each variance estimation method, the variance estimates were averaged across the Monte Carlo replicates. The ratio of the square root of this average of the estimated variances to the "actual" SD was computed for each variance estimation method. Boxplots of this ratio were constructed across all the simulation runs, by covariate, by number of clusters, and by the frailty parameter α , to illustrate how well the adjusted, bootstrap, and unadjusted SE estimates perform. These are shown in Figures 2 through 4 for covariates Z_1 through Z_3 .

One can see in Figures 2-4 that the ratios for the SE adjusted for clustering are quite close to 1 for all three covariates and all three levels of association α . The outliers in the box plots generally correspond to the smallest number of clusters ($m = 50$), the lowest proportion of type I events (25%), and the highest censoring percentage (50%). The ratios generally tend to be slightly below 1, yielding a slight underestimate of the variance. It is unclear why this occurs, but possibly it is because the estimating equations are only asymptotically unbiased. As the cluster size goes up, the ratios get closer to 1, and the estimated SE is quite accurate.

If this underestimation of the variance is a concern, bootstrap resampling of the clusters can be used to obtain a conservative (over) estimate of the SE. However, while it is conservative, the bootstrap estimate is not particularly accurate unless the number of clusters is large, in which case the sandwich estimate accounting for clustering also works well. The bootstrap variance estimate is also more difficult to implement since it is not available in standard software.

In contrast, Figures 2-4 illustrate that the SE estimate unadjusted for clustering is inconsistent when there is correlation present within clusters ($\alpha = 0.25, 0.5$), except in study designs where the covariate is completely independent of the cluster (as in Figure 3). In the matched pair type design (Figure 2), where each cluster has representation from both levels of the covariate Z_1 , the SE unadjusted for clustering is too high since it does not reflect the within cluster difference information. In contrast, when all individuals within a cluster have the same level of the covariate Z_3 (Figure 4), the SE unadjusted for clustering is too low, indicating that assuming independence overstates the information about the covariate.

4. Example

We illustrate this methodology on a retrospective cohort dataset from the registry of the Center for International Blood and Marrow Transplant Research (CIBMTR). This study investigates risk factors for transplant-related mortality (TRM) among adult patients age 18-50 with Acute Myelogenous Leukemia (AML) receiving a myeloablative hematopoietic stem cell transplant (HSCT) from an unrelated donor in the years 1995-2004. The outcome TRM is defined as death in continuous complete remission; relapse of disease is considered the competing risk. The study population contained 1314 patients transplanted at 151 transplant centers. Among these patients there were 557 TRM events, 374 relapses, and 383 censored individuals. There is a significant center effect ($p < 0.001$) on the cause-specific hazard of TRM, using the random effect score test of Commenges and Andersen (1995). Here we consider the impact of accounting for the clustering effect of patients transplanted at the same center on the standard errors of the regression coefficients.

Of particular interest in this study is to consider the effect of treatment-related factors on the cumulative incidence of TRM, while adjusting for patient- and donor- related factors. These treatment-related factors include the conditioning regimen (Busulfan + Cyclophosphamide (BuCy) vs. Cyclophosphamide + Total Body Irradiation (CyTBI)), the graft type (Peripheral Blood (PB) vs. Bone Marrow (BM)), and the use of graft-versus-host disease prophylaxis (CSA+MTX vs. CSA (No MTX) vs. FK-506 based vs. T-cell depletion). Patient- and donor-related factors adjusted for in the regression model include age at transplant, Karnofsky Performance Score, presence of coexisting disease, and the quality of the HLA match between donor and recipient.

Note that treatment-related factors are often part of a comprehensive center practice, so that patients at the same center are more likely to get the same conditioning regimen, for example. This would indicate that if the patients are correlated within center, ignoring the center effect in the calculation of the standard error would tend to underestimate the standard error corresponding to these factors.

Marginal models were fit to the TRM data using the pseudo-value method described in this paper. A description of the SAS code used to implement the methods is provided in Web Appendix A. Five time points were used, corresponding to the 10th, 30th, 50th, 70th, and 90th percentiles of the observed event times. A logistic link function was used to model the cumulative incidence probability as a function of the covariates. Standard errors were computed both ignoring the clustering effect and adjusting for the cluster unit defined by the transplant center, with an independent working correlation matrix. The results are given in Table 1.

While the parameter estimates for most of the patient characteristics have similar standard errors whether or not the center effect is accounted for, the standard errors for the treatment-related factors are underestimated when the center effect is ignored. This is likely because the centers tend to have consistent treatment regimens across patients. In this example, the magnitude of the underestimation is not large and does not lead to a qualitative change in the findings. This indicates that the within center correlation is more modest than that studied in the simulations.

5. Discussion

We have described a method for modeling the marginal cumulative incidence curve as a function of covariates, when the data are clustered. Generalized estimating equations and a sandwich variance estimate are used to adjust the standard errors for the clustering. This adjustment can have a substantial impact on the standard errors of the parameter estimates,

depending on the degree of within cluster association and the relationship between the covariate and the cluster. Simulations and examples were illustrated with an independence working correlation matrix, which provides consistent but likely inefficient parameter estimates. Additional research is needed to investigate appropriate working correlation structures accounting for both the correlation between individuals within a cluster, as well as the correlation across timepoints within an individual.

Appendix A

Here we derive the representation of the pseudovalue in (2). Using the IPCW representation of $\hat{F}_1(t)$ in (1), the pseudovalue can be written as

$$Y_{ij}(t) = \sum_{k,l} \frac{N_{kl}(t)}{\widehat{G}(X_{kl})} - \sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)}{\widehat{G}^{(ij)}(X_{kl})},$$

where $\widehat{G}^{(ij)}(t)$ is the Kaplan-Meier estimate of the censoring distribution omitting observation j in cluster i , and the second summation is taken over all clusters k and within cluster observations l except for observation j in cluster i . Then

$$Y_{ij}(t) = \sum_{k,l} \frac{N_{kl}(t)}{G(X_{kl})} - \sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)}{G(X_{kl})} \quad (7)$$

$$+ \sum_{(k,l) \neq (i,j)} N_{kl}(t) \left[\left\{ \frac{1}{\widehat{G}(X_{kl})} - \frac{1}{G(X_{kl})} \right\} - \left\{ \frac{1}{\widehat{G}^{(ij)}(X_{kl})} - \frac{1}{G(X_{kl})} \right\} \right] \quad (8)$$

$$+ N_{ij}(t) \left\{ \frac{1}{\widehat{G}(X_{ij})} - \frac{1}{G(X_{ij})} \right\} \quad (9)$$

The term on the right hand side of (7) simplifies to $N_{ij}(t) / G(X_{ij})$. The term in (9) is of order $O_p(n^{-1/2})$. To simplify the term in (8), first define $M^c(u) = \sum_{i,j} M_{ij}^c(u)$ and $M^{c(ij)}(u) = \sum_{(k,l) \neq (i,j)} M_{kl}^c(u)$. Since

$$\frac{1}{\widehat{G}(X_{kl})} - \frac{1}{G(X_{kl})} = \frac{1}{\widehat{G}(X_{kl})} \int_0^{X_{kl}} \frac{\widehat{S}(u^-) J(u)}{S(u) R(u)} dM^c(u),$$

the term in (8) can be written as

$$\sum_{(k,l) \neq (i,j)} N_{kl}(t) \left\{ \frac{1}{\widehat{G}(X_{kl})} \int_0^{X_{kl}} \frac{\widehat{G}(u^-) J(u)}{G(u) R(u)} dM^c(u) - \frac{1}{\widehat{G}^{(ij)}(X_{kl})} \int_0^{X_{kl}} \frac{\widehat{G}^{(ij)}(u^-) J(u)}{G(u) R^{(ij)}(u)} dM^{c(ij)}(u) \right\}.$$

Since $M^c(u) = M^{c(ij)}(u) + M_{ij}^c(u)$, this can be rearranged as

$$\sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)}{\widehat{G}(X_{kl})\widehat{G}^{(ij)}(X_{kl})} \int_0^{X_{kl}} \left\{ \frac{\widehat{G}^{(ij)}(X_{kl})\widehat{G}(u^-)J(u)}{G(u)R(u)} - \frac{\widehat{G}(X_{kl})\widehat{G}^{(ij)}(u^-)J(u)}{G(u)R^{(ij)}(u)} \right\} dM^{c(ij)}(u) + \sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)}{\widehat{G}(X_{kl})} \int_0^{X_{kl}} \frac{\widehat{G}(u^-)J(u)}{G(u)R(u)} dM_{ij}^c(u). \quad (10)$$

Consider the first term in expression (10). Define $r(u)$ such that $R(u)/n \rightarrow r(u)$ in probability. Note that

$$\int_0^{X_{kl}} \left\{ \frac{(n-1)\widehat{G}^{(ij)}(X_{kl})\widehat{G}(u^-)J(u)}{G(u)R(u)/n} - \frac{n\widehat{G}(X_{kl})\widehat{G}^{(ij)}(u^-)J(u)}{G(u)R^{(ij)}(u)/(n-1)} \right\} \frac{dM^{c(ij)}(u)}{\sqrt{n}},$$

converges in distribution to a random variable W_{kl} (say), so that the first term in (10) is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)W_{kl}}{(n-1)\widehat{G}(X_{kl})\widehat{G}^{(ij)}(X_{kl})} = O_p(n^{-1/2}).$$

The final remaining term in (10) is asymptotically equivalent to

$$\sum_{(k,l) \neq (i,j)} \frac{N_{kl}(t)}{G(X_{kl})} \int_0^{X_{kl}} \frac{J(u)}{R(u)} dM_{ij}^c(u). \quad (11)$$

Since

$$\int_0^{X_{kl}} \frac{J(u)}{R(u)} dM_{ij}^c(u) = \int_0^{X_{ij}} \frac{J(u)I(u \leq X_{kl})}{R(u)} dM_{ij}^c(u),$$

the term in (11) is equivalent to

$$\int_0^{X_{ij}} \frac{J(u)dM_{ij}^c(u)}{R(u)/n} \sum_{(k,l) \neq (i,j)} \frac{dN_{kl}(t)I(X_{kl} \geq u)}{nG(X_{kl})},$$

which following (1) is asymptotically equivalent to

$$\int_0^{X_{ij}} \frac{P(T \leq t, \varepsilon=1|T \geq u)}{G(u)} dM_{ij}^c(u).$$

Combining these results yields (2).

Acknowledgments

This research was partially supported by a grant (R01 CA54706-10) from the National Cancer Institute.

References

- Andersen PK, Klein JP, Rosthøj S. Generalized linear models for correlated pseudo-observations with applications to multi-state models. *Biometrika*. 2003; 90:15–27.
- Chen BE, Kramer JL, Greene MH, Rosenberg PS. Competing risks analysis of correlated failure time data. *Biometrics*. 2007; 64:172–179. [PubMed: 17680835]
- Commenges D, Andersen PK. Score test of homogeneity for survival data. *Lifetime Data Analysis*. 1995; 1:145–156. [PubMed: 9385097]
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *JASA*. 1999; 94:496–509.
- Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*. 2009; 15:241–255. [PubMed: 19051013]
- Hougaard, P. *Analysis of Multivariate Survival Data*. Springer-Verlag; New York: 2000.
- Katsahian S, Resche-Rigon M, Chevret S, Porcher R. Analyzing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. *Statistics in Medicine*. 2006; 25:4267–4278. [PubMed: 16960919]
- Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudo-values of the cumulative incidence function. *Biometrics*. 2005; 61:223–229. [PubMed: 15737097]
- Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. *Comput Methods Programs Biomed*. 2008; 89:289–300. [PubMed: 18199521]
- Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. *Statistics in Medicine*. 2007; 26:4505–4519. [PubMed: 17348080]
- Lee, EW.; Wei, LJ.; Amato, DA. Cox-type Regression analysis for large numbers of small groups of correlated failure time observations. In: Klein, JP.; Goel, PK., editors. *Survival Analysis: State of the Art*. Dordrecht: Kluwer Academic; 1992. p. 237–247.
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 78:13–22.
- Scheike TH, Zhang M-J, Gerds TA. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*. 2008; 95:205–220.
- Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *JASA*. 1998; 93:1164–1175.
- Satten GA, Datta S. Inverse-Probability-of-Censoring Weighted Average. *American Statistician*. 2001; 55:207–210.

Figure 1. Boxplot of relative bias (Bias/β) for a partitioning of simulation settings, by number of clusters m , observations per cluster n_i , % type I events, and censoring percentage. Boxplots represent the range of values across all other simulation parameters, i.e. across α and the covariate number.

Figure 2.

Boxplot of ratio of square root of average variances of parameter estimates to the Monte Carlo estimated SD, for covariate Z_1 . Boxplots are for ratios across all simulation scenarios, and are shown separately by the variance estimation method (Unadjusted for clustering, Adjusted for clustering, and Bootstrap estimate), by number of clusters m , and across 3 panels for varying levels of association, α . A reference line at the target ratio of 1.0 is also provided.

Figure 3.

Boxplot of ratio of square root of average variances of parameter estimates to the Monte Carlo estimated SD, for covariate Z_2 . Boxplots are for ratios across all simulation scenarios, and are shown separately by the variance estimation method (Unadjusted for clustering, Adjusted for clustering, and Bootstrap estimate), by number of clusters m , and across 3 panels for varying levels of association, α . A reference line at the target ratio of 1.0 is also provided.

Figure 4.

Boxplot of ratio of square root of average variances of parameter estimates to the Monte Carlo estimated SD, for covariate Z_3 . Boxplots are for ratios across all simulation scenarios, and are shown separately by the variance estimation method (Unadjusted for clustering, Adjusted for clustering, and Bootstrap estimate), by number of clusters m , and across 3 panels for varying levels of association, α . A reference line at the target ratio of 1.0 is also provided.

Table 1

Estimates, Standard Errors, and p-values for TRM in the unrelated donor HSCT for AML study. Results are shown for method adjusting for cluster and not adjusting for cluster

Factor	Comparison	Adjusted		Unadjusted	
		β	SE	SE	p-value
Age	31-40 vs. 18-30	-0.002	0.138	0.146	0.989
	41-50 vs. 18-30	0.258	0.154	0.141	0.068
KPS	< 90 vs. $\geq$ 90	0.321	0.115	0.121	0.008
	Yes vs. No	0.368	0.118	0.125	0.003
Coexisting Dz	2000-2004 vs. 1995-1999	-0.307	0.148	0.135	0.023
Year	Partially Matched vs. Matched	0.050	0.145	0.140	0.723
	Mismatched vs. Matched	0.751	0.156	0.176	< 0.001
Graft Type	PB vs. BM	-0.588	0.183	0.166	< 0.001
Conditioning	BuCy vs. CyTBI	-0.096	0.155	0.125	0.441
	CSA (No MTX) vs. CSA+MTX	0.695	0.313	0.267	0.009
GVHD Prophylaxis	FK506 based vs. CSA+MTX	-0.083	0.199	0.155	0.593
	TCD vs. CSA+MTX	0.315	0.198	0.161	0.051