Lifetime Data Anal (2009) 15:241–255
DOI 10.1007/s10985-008-9107-z

On pseudo-values for regression analysis in competing risks models

Frederik Graw · Thomas A. Gerds · Martin Schumacher

Received: 14 March 2008 / Accepted: 5 November 2008 / Published online: 3 December 2008

Abstract For regression on state and transition probabilities in multi-state models Andersen et al. (Biometrika 90:15–27, 2003) propose a technique based on jackknife pseudo-values. In this article we analyze the pseudo-values suggested for competing risks models and prove some conjectures regarding their asymptotics (Klein and Andersen, Biometrics 61:223–229, 2005). The key is a second order von Mises expansion of the Aalen-Johansen estimator which yields an appropriate representation of the pseudo-values. The method is illustrated with data from a clinical study on total joint replacement. In the application we consider for comparison the estimates obtained with the Fine and Gray approach (J Am Stat Assoc 94:496–509, 1999) and also time-dependent solutions of pseudo-value regression equations.

Keywords Competing risks · Generalized estimating equation · Jackknife pseudo-values · Regression models · Survival analysis · Von Mises expansion

F. Graw (✉)
Institute of Integrative Biology, ETH Zurich, Universitätsstr. 16,
8092 Zurich, Switzerland
e-mail: frederik.graw@env.ethz.ch

F. Graw · M. Schumacher
Institute of Medical Biometry and Medical Informatics,
University Medical Center Freiburg, Stefan-Meier-Str. 26,
79104 Freiburg, Germany

T. A. Gerds
Department of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, 1014 Copenhagen K, Denmark

1 Introduction

Regression models for multi-state processes have a wide range of applications. In this article, we are interested in estimating effects of prognostic factors on disease stage probabilities. For illustration, a clinical study is considered of an elderly population of patients with a hip prosthesis. Here the aim is to analyze the effects of factors like stem type, age or gender on the probability of revision due to aseptic loosening of the prosthesis before death (Fig. 1).

A useful and popular tool to perform regression analysis in such a situation with in general $R \in \mathbb{N}$ competing risks is the proportional hazard model for the sub-distribution function of a competing risk (Fine and Gray 1999); it corresponds with the link function $g(x) = \log\{-\log(1-x)\}$ in the transformation model

$$F_r^*(t \mid Z_i) = g^{-1}\left\{\beta_r^0(t) + \sum_{j=1}^{p} \beta_r^j Z_i^j\right\} = g^{-1}(\beta_{r,t}^T Z_i). \tag{1}$$

Here $\beta_r^0(t)$ is a baseline risk function, $\beta_{r,t} = \{\beta_r^0(t), \beta_r^1, \ldots, \beta_r^p\}^T$ and F_r^* the conditional cumulative incidence function corresponding to cause $r \in \{1, \ldots, R\}$, i.e. $F_r^*(t \mid Z_i)$ is the conditional probability that the event of interest occurs to patient i until time point t given covariates. The statistical analysis via generalized linear modeling is not straightforward because in practice, due to right censoring, the event times are often not observable for all patients. To overcome difficulties Andersen et al. (2003) proposed to use pseudo-values derived from the Aalen-Johansen estimate of the cumulative incidence function. To fit the model (1) the pseudo-values are evaluated at a fixed number of time points and then used as response in a generalized estimating equation (*GEE*) approach (Liang and Zeger 1986). Benchmarking the so obtained estimates for the regression coefficients to those of the familiar Fine and Gray (1999) approach yielded promising results in simulated and in real data (Andersen et al. 2003; Klein and Andersen 2005). However, a theoretical justification and a thorough investigation of the requirements under which the pseudo-value approach works seems to be lacking.

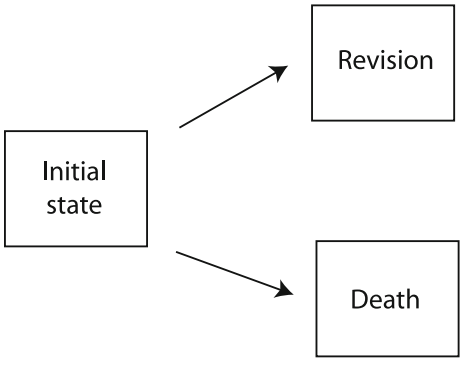

Fig. 1 Competing risks model for the time course of patient with hip prosthesis regarding two competing subsequent states

In this article we prove some conjectures of Andersen et al. (2003) and thereby justify the pseudo-value regression approach. We limit our investigation to regression models for the cumulative incidence function in a competing risks model. However, it seems that some of the arguments can be extended to yield similar results for regression on state occupation probabilities in more general multi-state models.

In order to be able to use an appropriate theorem of the *GEE* approach we have to relax "unbiasedness of the pseudo-values", as it is formulated in Andersen et al. (2003), to "conditional unbiasedness of the pseudo-values given the covariates". We show that the pseudo-values derived from the Aalen-Johansen estimate satisfy this new condition when the censoring mechanism is independent of the covariates and of the event times. The central argument is a second order von Mises expansion of the Aalen-Johansen estimate which leads to an appropriate representation of the jack-knife pseudo-values. Based on this representation we discuss the relation between the pseudo-value approach and the closely related approach of Scheike and Zhang (2007) and Scheike et al. (2008).

In Sect. 6 we apply the pseudo-value approach to data from a clinical study on total joint replacement (Maurer et al. 2001). In particular, we examine time-dependent solutions of the pseudo-value regression model and compare them to the time-constant regression coefficients obtained with the Fine and Gray (1999) approach.

2 Definitions and requirements

2.1 Competing risks data

Consider data from the time courses of patients in a competing risks model like the one depicted in Fig. 1. At a common time-origin each patient $i \in \{1, \ldots, n\}$ is in the initial state and a p-dimensional vector of covariates Z_i is recorded. At the event time T_i the course ends in one of the states $D_i \in \{1, \ldots, R\}$ representing a competing risk. We introduce the counting process $N_{ir}(t) = \mathcal{I}(T_i \leq t, D_i = r)$ whose expected value is the marginal cumulative incidence function of cause r. This can also be written as the expectation of the conditional probability to experience risk r before time t given covariates, with the expectation referring to the distribution of Z_i:

$$F_r(t) = E\{N_{ir}(t)\} = E[E\{N_{ir}(t) \mid Z_i\}] = E\{F_r^*(t \mid Z_i)\}.$$

The transition time T_i is right censored at the last time where the patient was observed to be in the initial state. For inference from censored data it is often required that the censoring mechanism satisfies some independence condition. In the approach of Andersen et al. (2003) it is assumed that:

$$\text{The censoring time } C_i \text{ is stochastically independent of } (T_i, D_i, Z_i). \qquad \text{(A1)}$$

The observed data are $X_i = (\tilde{T}_i, \Delta_i, D_i, Z_i)$ where $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = \mathcal{I}(T_i \leq C_i)$. Based on the observed data we also define the counting process $Y_i(t) = \mathcal{I}(\tilde{T}_i > t)$. Under (A1) the expectation of Y_i is the product of the marginal event free survival function $S(t)$ and the survival function of the censoring time $G(t)$:

$$E\{Y_i(t)\} = P(\tilde{T}_i > t) = P(T_i > t)\, P(C_i > t) = S(t)\, G(t).$$

The expectation of the counting process $\tilde{N}_{ir}(t) = \mathcal{I}(\tilde{T}_i \le t, \Delta_i = 1, D_i = r)$ calculated at $\mathrm{d}t$ also factorizes under (A1):

$$E\{\tilde{N}_{ir}(\mathrm{d}t)\} = E\{\Delta_i\, N_{ir}(\mathrm{d}t)\} = G(t-)\, F_r(\mathrm{d}t) = \tilde{F}_r(\mathrm{d}t). \tag{2}$$

For the theoretical results to hold, another important condition is the following:

$$\text{Consider only time points } t < \tau \text{ such that } G(\tau) > \nu > 0. \tag{A2}$$

Under (A2) we can rewrite the marginal cumulative incidence function by using (2) as

$$F_r(t) = \int_0^t \frac{G(s-)}{G(s-)} F_r(\mathrm{d}s) = \int_0^t \frac{\tilde{F}_r(\mathrm{d}s)}{G(s-)}. \tag{3}$$

Next we need the product limit form of the survival function of the censoring times. Introduce the distribution function $H(t) = P(\tilde{T}_i > t)$, the sub-distribution function $H^0(t) = P(\tilde{T}_i \le t, \Delta_i = 0)$ and the following empirical estimates $H_n(t) = n^{-1} \sum_i Y_i(t)$ and $H_n^0(t) = n^{-1} \sum_i \mathcal{I}(\tilde{T}_i \le t, \Delta_i = 0)$. Using these notations, the function G and its Kaplan–Meier estimate can be expressed as

$$G(t) = \prod_{s=0}^{t} \left\{ 1 - \frac{H^0(\mathrm{d}s)}{H(s-)} \right\} \quad \text{and} \quad \hat{G}(t) = \prod_{s=0}^{t} \left\{ 1 - \frac{H_n^0(\mathrm{d}s)}{H_n(s-)} \right\},$$

respectively. The inverse of the probability of censoring weighted (IPCW) estimate of F_r (Satten and Datta 2001; Jewell et al. 2007) is motivated by (3) and given by

$$\hat{F}_r(t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^s \frac{\tilde{N}_{ir}(\mathrm{d}s)}{\hat{G}(s-)}.$$

The estimate is due to Aalen and Johansen (1978) and traditionally it is obtained with the product limit form for transition matrices. Note that if there are ties between the event and the censoring times then $H_n(t)$ used in $\hat{G}(t)$ should be replaced by $n^{-1} \sum_i Y_i(t) - n^{-1} \sum_r \sum_i \tilde{N}_{ir}(t)$ (Satten and Datta 2001).

2.2 Jackknife pseudo-values

Let $\hat{G}^{(k)}$ denote the Kaplan–Meier estimate for G when it is computed based on the reduced sample $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n$. Correspondingly define the Aalen-Johansen estimate without the kth observation:

$$\hat{F}_r^{(k)}(t) = \frac{1}{(n-1)} \sum_{i \ne k} \int_0^s \frac{\tilde{N}_{ir}(\mathrm{d}s)}{\hat{G}^{(k)}(s-)}.$$

The idea of Klein and Andersen (2005) is to use the jackknife pseudo-values of the Aalen-Johansen estimate given by

$$J_{kr}(t) = n\,\hat{F}_r(t) - (n-1)\,\hat{F}_r^{(k)}(t) \tag{4}$$

as response in a *GEE* approach. Note that these pseudo-values are exactly equal to $N_{kr}(t)$ in the special case where all observations are uncensored, since then $\hat{G}(t) = 1$. Thus, in the uncensored situation $E\{J_{kr}(t)\} = E\{N_{kr}(t)\} = F_r(t)$ and also

$$E\{J_{kr}(t) \mid Z_k\} = F_r^*(t \mid Z_k) \tag{5}$$

at all time points t.

2.3 Generalized estimating equations

For estimating the parameters of the model (1) Klein and Andersen (2005) propose to select a finite number of time points $t_1, \ldots, t_l$ and to evaluate a vector of pseudo-values to be used as (pseudo) responses for patient k: $\{J_{kr}(t_1), \ldots, J_{kr}(t_l)\}^T$. In the sequel, for notational convenience and without loosing substance of the mathematical problem, we only consider a single time point t and work with the pseudo response $J_{kr}(t)$. Extensions to multiple time points are discussed in Sect. 5.

In the pseudo-value approach estimates of the regression coefficients are the solutions of the following generalized estimating equation:

$$U_{(n)}(\beta_{r,t}) = \sum_{k=1}^{n} \left(\frac{\partial g^{-1}(\beta_{r,t}^T Z_k)}{\partial \beta_{r,t}} \right)^T V_k^{-1} \left\{ J_{kr}(t) - g^{-1}(\beta_{r,t}^T Z_k) \right\} = 0. \tag{6}$$

Here V_k is the usual "working covariance matrix" which may account for the correlation structure inherent to the pseudo response values (Andersen et al. 2003). In case of a single time point, we have $V_k = 1$. It seems straightforward to use a general theorem for *GEE* (Liang and Zeger 1986) to prove the large sample properties of the solution $\hat{\beta}_{r,t}$ to Eq. 6. However, in addition to the usual regularity conditions the following "asymptotic unbiasedness" of the pseudo-values is required:

$$E\{J_{kr}(t) \mid Z_k\} = g^{-1}(\beta_{r,t}^T Z_k) + o_P(1). \tag{7}$$

This holds trivially and without remainder term if all event times are uncensored, as outlined before in Eq. 5. It is not straightforward and surprisingly difficult to verify this condition of the pseudo-values for right censored situation. Condition (7) is needed to show that $E\{U_{(n)}(\beta_{r,t}^*)\} = 0$ at the true parameter value $\beta_{r,t}^*$. In Andersen et al. (2003) it is argued that unbiasedness of (6) follows directly from their equation (2.6) which translated to our setting states that $E\{J_{kr}(t)\} = E\{g^{-1}(\beta_{r,t}^T Z_k)\}$. This however seems to be valid only in the special case where $\beta_{r,t}^j = 0$ for $j \geq 1$.

3 Von Mises expansion

In this section we obtain a representation of the jackknife pseudo-values for the competing risks model given in (4). The representation is then used to show validity of (7). It also allows a comparison of the jackknife pseudo-values and the weighted binomial response considered in Scheike et al. (2008). We discuss this further in Sect. 5.

As a first step we define the Aalen-Johansen functional. Let P denote the probability law of the vector X_i and $P_n(\cdot) = n^{-1} \sum_i \mathcal{I}(X_i \in \cdot)$, $i = 1, \ldots, n$ the empirical law corresponding to the sample of right censored observations $X_1, \ldots, X_n$. Further denote by $P_n^{(k)}$ the empirical distribution of the reduced sample $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n$. The Aalen-Johansen functional $\psi_r : \mathcal{P} \to \mathcal{F}$ operates on a set $\mathcal{P}$ of probability measures for X_i which includes P and the empirical measures (James 1997; Gill 1989) and maps into the set $\mathcal{F}$ of all sub-distribution functions. It is defined such that $\psi_r(P) = F_r$ is the parameter of interest and $\psi_r(P_n) = \hat{F}_r$ the Aalen-Johansen estimate corresponding to the sample $X_1, \ldots, X_n$. Hence, the jackknife pseudo-values can be expressed as $J_{kr} = n\, \psi_r(P_n) - (n-1)\, \psi_r(P_n^{(k)})$.

Generally, a smooth statistical functional ψ can be "von Mises expanded" in a similar way as a smooth function can be "Taylor expanded" (Gill 1989):

$$\psi(P_n) = \psi(P) + n^{-1} \sum_{i=1}^{n} \dot{\psi}(X_i) + \frac{1}{2} n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \ddot{\psi}(X_i, X_j) + \mathrm{O}_P(n^{-\frac{3}{2}}). \quad (8)$$

Here $\dot{\psi}$ and $\ddot{\psi}$ are the first and second order Gateaux derivatives of the functional ψ, also called influence functions (Hampel 1974) or "canonical gradients". The first derivative is centered, $E\{\dot{\psi}(X_i)\} = 0$ (Huber 1977); the second is symmetric, $\ddot{\psi}(X_i, X_j) = \ddot{\psi}(X_j, X_i)$ and ought to satisfy for every y (see Van der Vaart 1998, Sect. 20.1.1)

$$E\{\ddot{\psi}(X_i, y)\} = \int \ddot{\psi}(x, y) \mathrm{d}P(x) = 0. \quad (9)$$

Theorem 1 *For a differentiable statistical functional ψ, which possesses a second order von Mises expansion as in (8) such that also (9) holds, the jackknife pseudo-values are represented by:*

$$n\, \psi(P_n) - (n-1)\, \psi(P_n^{(k)}) = \psi(P) + \dot{\psi}(X_k) + \mathrm{o}_P(1).$$

Proof The representation follows from Eq. 8:

$$\begin{aligned} & n\, \psi(P_n) - (n-1)\psi(P_n^{(k)}) \\ & \quad = n\, \psi(P) - (n-1)\psi(P) + \sum_{i=1}^{n} \dot{\psi}(X_i) - \sum_{i \neq k} \dot{\psi}(X_i) \\ & \qquad + \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{n} \ddot{\psi}(X_i, X_j) - \frac{1}{2(n-1)} \sum_{i \neq k} \sum_{j \neq k} \ddot{\psi}(X_i, X_j) \end{aligned}$$

$$
\begin{aligned}
&+ n\,\mathrm{O}_P(n^{-\frac{3}{2}}) - (n-1)\,\mathrm{O}_P((n-1)^{-\frac{3}{2}}) \\
&= \psi(P) + \dot{\psi}(X_k) \\
&\quad - \frac{1}{2n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} \ddot{\psi}(X_i, X_j) + (n-1)^{-1} \sum_{i=1}^{n} \ddot{\psi}(X_i, X_k) \\
&\quad - \frac{1}{2(n-1)} \ddot{\psi}(X_k, X_k) + n^{-\frac{1}{2}}\mathrm{O}_P(1) - (n-1)^{-\frac{1}{2}}\mathrm{O}_P(1) \\
&= \psi(P) + \dot{\psi}(X_k) + (n-1)^{-1} \sum_{i=1}^{n} \ddot{\psi}(X_i, X_k) + \mathrm{o}_P(1).
\end{aligned}
$$

By the law of large numbers, the third term in the last line of the previous display converges to

$$E\{\ddot{\psi}(X_i, y)\} = \mathrm{o}_P(1)$$

by Eq. 9. □

The Aalen-Johansen functional is Hadamard differentiable (see e.g. Gill and Johansen 1990) . It is also well-known that the Aalen-Johansen estimate is $n^{\frac{1}{2}}$-consistent (Andersen et al. 1993). Thus, by arguing similar as in James (1997) who considered the Kaplan–Meier and the Nelson-Aalen estimator, one may show that ψ_r has a second order von Mises expansion as in (8). The crucial assumption for this is our condition (A2). Indeed, it has been shown in Jewell et al. (2007, Appendix A) that the (first order) influence curve of the Aalen-Johansen estimate $\psi_r(P_n)$ for the cumulative incidence function in the competing risks model is given by

$$\dot{\psi}_r(X_i;\, P) = \frac{\tilde{N}_{ir}(t) - \Delta_i\, F_r(t)}{G(T_i-)} + \int_0^{\tilde{T}_i} \frac{P(T_i \le t,\, D_i = r \mid T_i \ge u) - F_r(t)}{G(u)}\, dM_G(u) \tag{10}$$

where $dM_G(u) = \mathcal{I}(\tilde{T}_i \in du, \Delta_i = 0) - \mathcal{I}(\tilde{T}_i \ge u)\Lambda_G(du)$ is the martingal and $\mathrm{d}\Lambda_G = -\mathrm{d}G/G$ the cumulative hazard function associated with $1 - G$. Thus, we have the following lemma to Theorem 1:

Lemma 1 *Under conditions* (A1) *and* (A2) *the following representation holds*

$$J_{kr}(t) = \dot{\psi}_r(X_k) + F_r(t) + \mathrm{o}_P(1)$$

Proof It remains to show that $(n-1)^{-1}\sum_i \ddot{\psi}_r(X_i, X_k) = \mathrm{o}_P(1)$. This follows directly from the fact that the second order influence functions of the Nelson-Aalen and the Kaplan–Meier functionals have this property (James 1997, see Sect. 5) □

3.1 Properties of the pseudo-values

The representation given in Lemma 1 immediately leads to some important properties of the jackknife pseudo-values. They are summarized in the following

Lemma 2 *Under conditions* (A1) *and* (A2) *we have:*

(i) $J_{kr}(t)$, $k = 1, \ldots, n$ *can be approximated by independent and identically distributed variables*
(ii) $E\{J_{kr}(t)\} = F_r(t) + o_P(1)$, $\forall k \in \{1, \ldots, n\}$
(ii) $E\{J_{kr}(t) \mid Z_k\} = F_r^*(t \mid Z_k) + o_P(1)$, $\forall k \in \{1, \ldots, n\}$.

Proof (i) and *(ii)* follow directly from the representation given in Lemma 1 and the properties of the (first order) influence function. For *(iii)* note first that under (A1)

$$\begin{aligned} E\{\mathrm{d}M_G(t) \mid Z_i\} &= E\{\mathcal{I}(C_i \in dt)(1-\Delta_i) \mid Z_i\} \\ &\quad + \frac{\mathrm{d}G(t)}{G(t)} E\{\mathcal{I}(T_i \geq t)\mathcal{I}(C_i > t) \mid Z_i\} \\ &= -S^*(t- \mid Z_i)\mathrm{d}G(t) + S^*(t- \mid Z_i)\mathrm{d}G(t)\frac{G(t)}{G(t)} = 0. \end{aligned}$$

Here $S^*(\cdot \mid Z_i)$ is the conditional event free survival function given Z_i. Thus, using also Eq. 2 and Lemma 1, we have under (A1) and (A2) for every k:

$$\begin{aligned} E\{J_{kr}(t) \mid Z_k\} &= F_r(t) + \frac{E\{\tilde{N}_{kr}(t) \mid Z_k\}}{G(T_k-)} - F_r(t)\frac{E\{\Delta_k \mid Z_k\}}{G(T_k-)} + o_P(1) \\ &= F_r(t) + F_r^*(t \mid Z_k) - F_r(t)\frac{G(T_k-)}{G(T_k-)} + o_P(1) \\ &= F_r^*(t \mid Z_k) + o_P(1) \end{aligned}$$

□

4 Asymptotics for pseudo-value estimation equations

We are now prepared for a rigorous investigation of consistency and asymptotic normality of the estimators defined by (6).

Theorem 2 *Consider a time point t that satisfies* (A2). *Under* (A1) *and mild regularity conditions regarding the link function $g(\cdot)$, the solution $\hat{\beta}_{r,t}$ to* (6) *is consistent and asymptotically normal for estimating the parameter $\beta_{r,t}^*$ of model* (1)*:*

$$\sqrt{n}(\hat{\beta}_{r,t} - \beta_{r,t}^*) \sim N(0, \Sigma_{r,t})$$

where the asymptotic variance $\Sigma_{r,t}$ is consistently estimated by the sandwich-form:

$$\hat{\Sigma}_{r,t} = \hat{\Gamma}_{r,t}^{-1}(\hat{\beta}_{r,t}) Var\left\{U_{(n)}(\hat{\beta}_{r,t})\right\} \hat{\Gamma}_{r,t}^{-1}(\hat{\beta}_{r,t})$$

where

$$\hat{\Gamma}_{r,t}(\hat{\beta}_{r,t}) = n^{-1} \sum_{k=1}^{n} \left\{ \frac{\partial g^{-1}(\beta_{r,t}^T Z_k)}{\partial \beta_{r,t}} \right\}^T V_k^{-1} \left\{ \frac{\partial g^{-1}(\beta_{r,t}^T Z_k)}{\partial \beta_{r,t}} \right\}$$

$$Var\left\{ U_{(n)}(\hat{\beta}_{r,t}) \right\} = n^{-1} \sum_{k=1}^{n} U_k(\hat{\beta}_{r,t}) U_k^T(\hat{\beta}_{r,t})$$

and $U_k(\cdot)$ *is denoted by Eq. 6 with* $U_{(n)}(\cdot) = \sum_k U_k(\cdot)$

Proof The essential part of the proof of consistency is to show that the residual

$$J_{kr}(t) - g^{-1}(\beta_{r,t}^T Z_k)$$

has expectation zero. This follows directly from Lemma 2 (iii). The essential part of the proof of asymptotic normality is to show that the score process can be approximated by a sum of independent and identically distributed random variables at the $n^{-\frac{1}{2}}$-rate. Such a representation follows directly from Lemma 1:

$$U_{(n)}(\beta_{r,t}) = \sum_{k=1}^{n} \left(\frac{\partial g^{-1}(\beta_{r,t}^T Z_k)}{\partial \beta_{r,t}} \right)^T V_k^{-1} \{ \dot{\psi}_r(X_k) - g^{-1}(\beta_{r,t}^T Z_k) \} + o_P(n^{-\frac{1}{2}}).$$

The rest of the proof is analogous to Liang and Zeger (1986). □

5 Extensions

5.1 Multiple time points

A single time point is enough to identify the regression coefficients of model (1). However, in practice it might be more efficient to use several or all time points. Furthermore, it might be of interest to estimate the baseline risk function $\beta_r^0(t)$. The score function (6) can be modified to fit model (1) based on a fixed set of time points (Andersen and Klein 2007) and based on all time points (Scheike and Zhang 2007). Mathematically the latter is more involved because then the model involves a nonparametric component (Scheike et al. 2008). In both approaches additional correlation structure is introduced through multiple jackknife pseudo-values from the same patient which is considered in the working covariance matrix V_k in (6) (see also Liang and Zeger 1986) .

5.2 Comparison with direct binomial regression

The jackknife pseudo-value approach discussed here is closely related to the direct binomial regression approach as given by Scheike et al. (2008). There the authors consider the weighted event indicator

$$\tilde{N}_{kr}(t)/\hat{G}^*(T_k \mid Z_k) \tag{11}$$

as response in a generalized linear model approach for patient k at time point t. Here $\hat{G}^*(t \mid Z_k) = P(C_k > t \mid Z_k)$ is an estimate of the conditional censoring survival function. Under (A1) the expression in (11) can be approximated by

$$\tilde{N}_{kr}(t)/G(T_k)$$

and this equals the first term of the influence function $\dot{\psi}_r$ given in (10). Thus, under (A1) the difference between the jackknife pseudo-value response $J_{kr}(t)$ and (11) is

$$\frac{\Delta_k F_r(t)}{G(T_k-)} + \int_0^{\tilde{T}_k} \frac{P(T_k \le t, D_k = r \mid T_k \ge u) - F_r(t)}{G(u)} dM_G(u) + o_P(1).$$

which could be interpreted as the "pseudo-part" considered in the first approach.

5.3 More general multi-state models

Andersen et al. (2003) formulated their pseudo-value approach for state occupation probabilities in general multi-state models. So far, we were able to analyze the validity of their propositions in the situation of competing risks. The claim of theorem 2 would be valid in case of general multi-state models as long as the pseudo-values satisfy a conditional unbiasedness condition as given in (7). Analogously, if $\theta(t)$ defines a state probability in a multi-state model, $\hat{\theta}(t)$ a corresponding consistent estimator based on random variables $X_1, \ldots, X_n$ and $J_{k,\theta}(t) = n\hat{\theta}(t) - (n-1)\hat{\theta}^{(k)}(t)$ the jackknife pseudo-value for $k \in \{1, \ldots, n\}$ analogous to (4), we would require

$$E\left\{J_{k,\theta}(t) \mid Z_k\right\} = g^{-1}(\beta_{r,t}^T Z_k) + o_P(1).$$

To proof this condition one might be able to find a similar representation for the jackknife pseudo-value as given in lemma 1. This might be possible as the arguments used can be adapted to general multi-state situations if we are able to construct $\hat{\theta}(t)$ by smooth mappings of the Nelson-Aalen and the Kaplan–Meier functionals.

6 Performing pseudo-value regression

In this section we apply the pseudo-value regression approach to data from a clinical study (Maurer et al. 2001) and compare the estimates to the results of the Fine and Gray (1999) approach. We extend the ideas of Andersen et al. (2003) by considering also time-dependent coefficients.

During the years 1987–1993 a total of 442 patients with newly implanted hip prostheses where followed concerning the two competing risks *death* and revision of the prosthesis (*rev*) due to aseptic loosening. Patients with two prostheses were assumed as *rev* when the first prosthesis failed. During the study-period ($t \in [0, 4474]$ days)

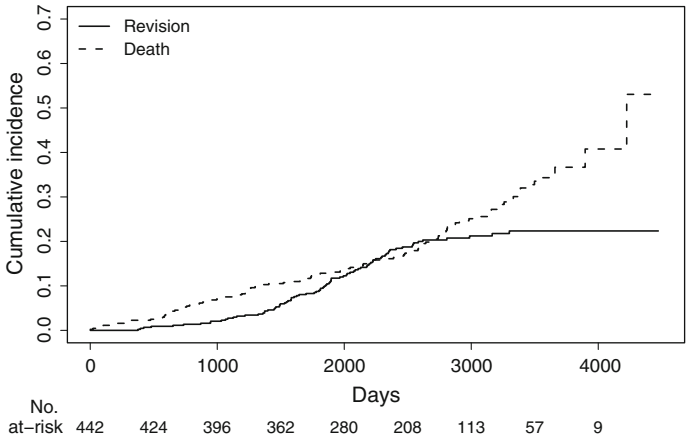

Fig. 2 Cumulative incidence functions $F_{\text{rev}}(t)$ and $F_{\text{death}}(t)$ for the data of Maurer et al. (2001) estimated by the Aalen-Johansen estimator

a total of $rev = 84$ and $death = 112$ were observed; 246 patients are right-censored. The cumulative incidence functions estimated by the Aalen-Johansen estimator are illustrated in Fig. 2. Several covariates were measured at time of implantation. We focused our investigations on the following variables which are either categorized or binary: *Stem type* with type *Titan SLS* or *Titan GS*, *Stem size* of the prosthesis in *mm* (7.5, 10, ≥12.5), *gender* and *age* (<65, 65–75, ≥75 years). They were found to be relevant in the previous analysis (Schwarzer et al. 2001). Analysis was performed in R (R Development Core Team 2006; Halekoh and Højsgaard 2006).

To investigate time-dependency the pseudo-values for each patient were calculated every 100 days. The model was fitted at each time point separately. Solutions for the covariates *Stem type* and *gender* are given in Fig. 3 with the corresponding pointwise 95% confidence intervals. The variation of these estimates is related to the Aalen Johansen estimates of the corresponding cumulative incidence functions (Fig. 2) as these are used to calculate the pseudo-values. As stated above, Klein and Andersen (2005) propose to consider several grid points simultaneously in the generalized estimating Eq. 6. Based on simulations they suggested to study at least 10 grid points which are equidistantly distributed on the scale of $\tilde{T}_i$ which would lead to stable estimates for β_r^j. Although we could confirm their results in our data set we assume that this number of grid points could be insufficient as the appropriate number might depend on the data structure.

By studying several time points simultaneously, the working covariance matrix V_i in (6) allows to consider different correlation structures taking into account the autocorrelation between the calculated pseudo-values for patient i. Klein and Andersen (2005) suggest to use an independent correlation structure as they find no advantage for extended versions. With regard to this observation and neglecting risk type r in the notation, we propose to use $\overline{\beta^j} = 1/t_{max} \int_0^{t_{max}} \hat{\beta}_s^j \, ds$ as the mean over $\hat{\beta}_t^j$ in time, with $t_{max} = \max_i(\tilde{T}_i)$. Estimates are compared in Table 1 and shown in Fig. 3. The

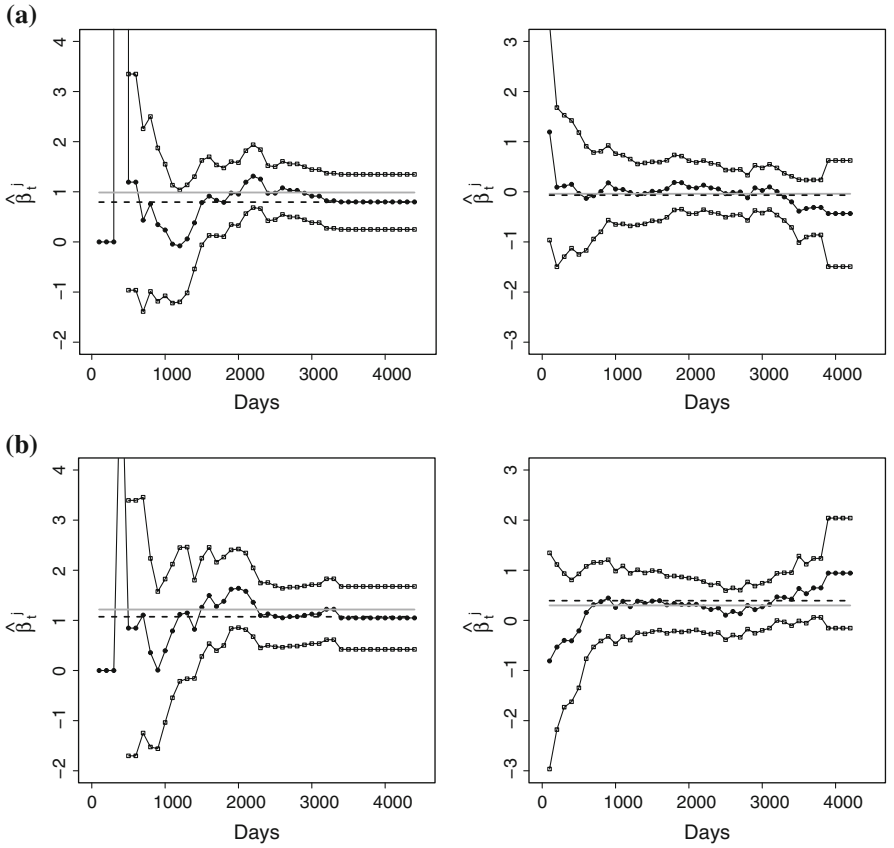

Fig. 3 Estimated regression parameters for $F_{\text{rev}}(t)$ *(left)* and $F_{\text{death}}(t)$ *(right)* for binary covariates *Stem Type* (a) and *gender* (b). Graphics contain single time point estimates $\hat{\beta}_t^j$ (•) and their pointwise 95% confidence interval (□), the mean $\overline{\beta^j}$ *(dashed line)* and $\hat{\beta}_{FG}^j$ *(solid line, grey)* obtained by the Fine & Gray method. For $F_{\text{rev}}(t)$ 95% confidence intervals start at $t = 500$ for graphical clarity due to estimates of $\hat{\beta}_{400}^j = 3.5e^{11}$ (a) and $\hat{\beta}_{400}^j = 6.8$ (b), respectively

time independent estimate $\hat{\beta}_{FG}$ given by Fine and Gray (1999) is added in Fig. 3. Analysis of all variables (figures not shown) give evidence for the approximation of $\hat{\beta}_{FG}$ as a (weighted) mean $\overline{\beta_t^j}$, although this proposition has to be checked in detail.

In Fig. 4 the model based cumulative incidence functions are given for the different values of the covariable $Stem\,type$. Predicted curves based on estimates of the pseudo-value approach using $\overline{\beta^j}$ and the Fine & Gray method are compared to results obtained by the Aalen-Johansen estimator. Predictions based on the pseudo-value approach are calculated only for the covariable $stem\,type$ setting all other covariates to zero. Model curves based on $\overline{\beta^j}$ show the lowest values. Caused by the direct estimation of the intercept term $\beta^0(t)$ these curves are not strictly monotone increasing. Further technique is required to force monotony.

Table 1 Estimators for β^j given by (1) $\overline{\beta^j}$, mean based on all $\hat{\beta}_t^j$ with $t \geq 500$, (2) $\hat{\beta}_{sim}^j$ estimated as proposed by Klein and Andersen (2005) based on 10 grid points (equidistantly distributed on $\tilde{T}_i$) and (3) the Fine & Gray method $\hat{\beta}_{FG}^j$

Z^j	$\overline{\beta^j}$	SE	$\hat{\beta}_{sim}^j$	SE	$\hat{\beta}_{FG}^j$	SE
(a) *Revision*						
Stem type						
Titan SLS	0.794	0.33	0.886	0.42	0.986	0.24
Stem size						
7.5	1.566	0.36	1.578	0.24	1.700	0.35
10.0	0.865	0.31	0.927	0.33	1.006	0.31
Gender						
Male	1.072	0.31	1.126	0.27	1.217	0.29
Age						
65–75 years	−0.691	0.43	−0.730	0.24	−0.767	0.25
≥75 years	−1.524	0.52	−1.579	0.34	−1.835	0.39
(b) *Death*						
Stem type						
Titan SLS	−0.067	0.19	−0.194	0.32	−0.040	0.20
Stem size						
7.5	−0.426	0.28	−0.757	0.40	−0.476	0.29
10.0	−0.565	0.18	−0.638	0.22	−0.546	0.21
Gender						
Male	0.393	0.25	0.496	0.23	0.298	0.20
Age						
65–75 years	0.922	1.10	0.647	0.33	0.939	0.35
≥75 years	2.101	1.00	1.918	0.31	2.128	0.33

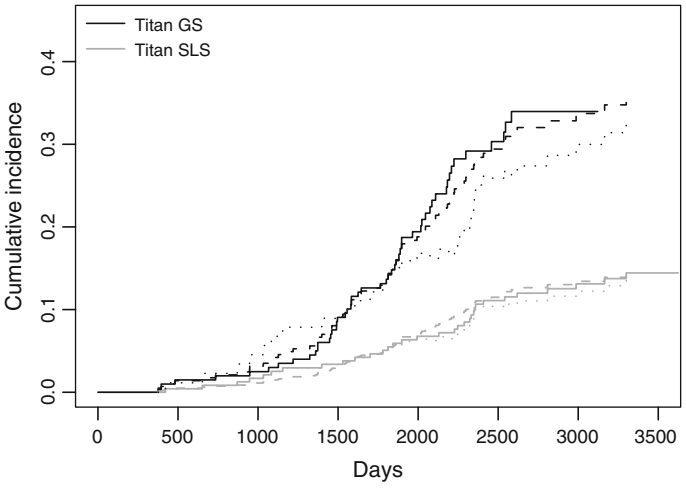

Fig. 4 Model based cumulative incidence functions $F_{\text{rev}}(t)$ for values of *Stem type*= $\{Titan\,SLS, Titan\,GS\}$. Predictions based on the Aalen-Johansen estimator *(solid line)*, the Fine & Gray method *(dashed)* and the pseudo-value approach with $\overline{\beta^j}$ *(dotted)*

7 Summary

In this work we examined the pseudo-value regression approach proposed by Andersen et al. (2003) in competing risks models. We have proven consistency and asymptotic normality of the estimates of regression coefficients for the cumulative incidence function. In addition to the usual mild regularity conditions needed for generalized estimating equations we have shown here that the jackknife pseudo-values derived from the Aalen-Johansen estimate of the cumulative incidence function satisfy a conditional unbiasedness condition. For this we used a higher order von Mises expansion of the Aalen-Johansen function. Similar expansions have been used earlier to analyze for example the variance of jackknifed estimates (Parr 1985).

The pseudo-value approach is a reasonable method for direct regression analysis on state and transition probabilities in multi-state models. In Scheike and Zhang (2007) a similar approach has been studied. In the illustration part we have considered the effect of the time at which the pseudo-values are calculated on the regression coefficients. We have shown that the time independent coefficients can be estimated by a natural extension of the jackknife pseudo-value approach. The proposed estimate is similar to the so far suggested extension (Andersen et al. 2003) which was based on a grid of about 10 time points. As single time point estimates suppose the choice of an appropriate grid may depend on the given data.

The pseudo-value approach requires independent censoring. This assumption might not always be satisfied. In practice, however, it is often reasonable to assume that the censoring mechanism is conditionally independent of the event times given the covariates. The IPCW approach can be modified to work under this weaker assumption (Scheike et al. 2008; Fine 1999).

So far, we have only considered competing risks models. The situation in the general multi-state model, especially if pseudo-values for state occupation probabilities in such circumstances satisfy the conditional unbiasedness condition, is not yet clear.

Acknowledgements We thank Per Kragh Andersen for introducing us to the problem and Jan Beyersmann for helpful discussions. We are grateful to Professor Peter Ochsner and Dr. Thomas Maurer for making the data of the hip prosthesis study available that has been conducted at the Kantonsspital Liestal, Switzerland. We thank all the anonymous referees, especially one in particular, for their helpful comments.

References

Aalen OO, Johansen S (1978) An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Stat 5:141–150

Andersen PK, Klein JP (2007) Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. Scand J Stat 34:3–16

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Andersen PK, Klein JP, Rosthøj S (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90(1):15–27

Fine JP (1999) Analysing competing risks data with transformation models. J R Stat Soc B 61(4):817–830

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94(446):496–509

Gill RD (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. Scand J Stat 16(2):97–128

Gill RD, Johansen S (1990) A survey of product-integration with a view toward application in survival analysis. Ann Stat 18(4):1501–1555

Halekoh U, Højsgaard S (2006) The R package geepack for generalized estimating equations. J Stat Softw 15(2):1–11

Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:1179–1186

Huber P (1977) Robust statistical procedures. SIAM, Philadelphia

James LF (1997) A study of a class of weighted bootstrap for censored data. Ann Stat 25(4):1595–1621

Jewell NP, Lei X, Ghani AC, Donnelly CA, Leung GM, Ho LM, Cowling BJ, Hedley AJ (2007) Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). Stat Med 26(9):1982–98

Klein JP, Andersen PK (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61(1):223–229

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13–22

Maurer TB, Ochsner PE, Schwarzer G, Schumacher M (2001) Increased loosening of cemented straight stem prostheses made from titanium alloys. An analysis and comparison with prostheses made of cobalt-chromium-nickel alloy. Int Orthop 25(2):77–80

Parr WC (1985) Jackknifing differentiable statistical functionals. J R Stat Soc B 47:56–66

R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Satten G, Datta S (2001) The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. Am Stat 55(3):207–210

Scheike T, Zhang MJ (2007) Direct modelling of regression effects for transition probabilities in multistate models. Scand J Stat 34(1):17–32

Scheike TH, Zhang MJ, Gerds TA (2008) Predicting cumulative incidence probability by direct binomial regression. Biometrika 95:205–220

Schwarzer G, Schumacher M, Maurer TB, Ochsner PE (2001) Statistical analysis of failure times in total joint replacement. J Clin Epidemiol 54(10):997–1003

Vander Vaart A (1998) Asymptotic statistics. Cambridge Univ. Press, Cambridge