Biometrika (2012), **99**, 1, *pp.* 167–184
doi: 10.1093/biomet/asr062
Advance Access publication 29 December 2011

Estimating treatment effects with treatment switching via semicompeting risks models: an application to a colorectal cancer study

BY DONGLIN ZENG

Department of Biostatistics, CB 7420, University of North Carolina, Chapel Hill, North Carolina 27516, U.S.A.

dzeng@bios.unc.edu

QINGXIA CHEN

Department of Biostatistics, Vanderbilt University, 1161 21st Avenue South, S-2323 Medical Center North, Nashville, Tennessee 37232, U.S.A.

cindy.chen@vanderbilt.edu

MING-HUI CHEN

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, Connecticut 06269, U.S.A.

ming-hui.chen@uconn.edu

JOSEPH G. IBRAHIM

Department of Biostatistics, CB 7420, University of North Carolina, Chapel Hill, North Carolina 27516, U.S.A.

ibrahim@bios.unc.edu

AND AMGEN RESEARCH GROUP

Amgen Inc., One Amgen Center Drive, Thousand Oaks, California 91320, U.S.A.

jpan@amgen.com

SUMMARY

Treatment switching is a frequent occurrence in clinical trials, where, during the course of the trial, patients who fail on the control treatment may change to the experimental treatment. Analysing the data without accounting for switching yields highly biased and inefficient estimates of the treatment effect. In this paper, we propose a novel class of semiparametric semicompeting risks transition survival models to accommodate treatment switches. Theoretical properties of the proposed model are examined and an efficient expectation-maximization algorithm is derived for obtaining the maximum likelihood estimates. Simulation studies are conducted to demonstrate the superiority of the model compared with the intent-to-treat analysis and other methods proposed in the literature. The proposed method is applied to data from a colorectal cancer clinical trial.

Some key words: Expectation-maximization algorithm; Maximum likelihood estimate; Noncompliance; Panitumumab; Partial switching; Transition model; Treatment switching.

1. INTRODUCTION

Treatment switching commonly occurs in clinical trials such as in cancer or in other diseases, where patients who fail on the control treatment may begin taking the experimental treatment. This often happens in cancer clinical trials when the control arm consists of a placebo or no treatment. In such trials, patients in the control arm who experience an intermediate event, such as disease progression, may begin taking the experimental treatment to receive a rescue medication. As discussed in Marcus & Gibbons (2001), an intent-to-treat analysis will lead to attenuated treatment effect estimates, and thus one must properly model the data accommodating this switching effect and then appropriately estimate the treatment effect.

In clinical trials, there are many types of switching possibilities. Drop-in refers to situations where control subjects start taking an active treatment. There is also switching due to drop-out, where subjects stop taking the active treatment. Here we focus on the drop-in problem of control subjects switching to the experimental treatment after experiencing an intermediate event.

Methods have been advocated to compensate for the effects of drop-in, assuming an intent-to-treat analysis. This could be a viable approach if the drug effect is believed to be sufficiently large to yield a clinically meaningful intent-to-treat effect. Various methods for sample size adjustment are described by Lachin & Foulkes (1986), Lakatos (1988), Lu & Pajak (2000), Porcher et al. (2002), Jiang et al. (2004) and Barthel et al. (2006). Although these approaches manage the risk of a false-negative error, they may result in a larger than needed sample size and yield an effect estimate of marginal clinical significance that fails to address the drop-in bias, especially when appreciable drop-in occurs nonrandomly. In the presence of drop-in, analysis methods to estimate the treatment's causal effect that do not respect randomization are potentially confounded. Common examples include an analysis treating the drop-in time as a censoring time, or the exclusion of patients with drop-in. Law & Kaldor (1996) proposed a multiplicative Cox model in which patients are divided into subgroups based on their randomized and observed subsequent therapy. As noted by White (1997), their model is flawed since subgroup membership at a particular time depends on the future, and estimates will tend to be biased under the null. Other approaches that respect the randomization include the use of causal models with counterfactuals (Lunceford et al., 2002; White et al., 2003; London et al., 2010). Related approaches are marginal structural models (Robins et al., 2000). For valid causal inference, these models must account for all confounders that predict drop-in and there must be no censoring bias. Even when these conditions apply, model estimates can become unstable when drop-in is certain among all patients with a specific value of a time-dependent covariate. Examples of applying marginal structural models are provided by Hernán et al. (2000) and Yamaguchi & Ohashi (2004).

When marginal structural models are not applicable, a structural nested model may be used (Yamaguchi & Ohashi, 2004; Greenland et al., 2008) In particular, based on the methods of Robins & Tsiatis (1991), Branson & Whitehead (2002) developed an estimation method for an accelerated failure-time model similar to a structural nested model to estimate the true effect. Their model assumes that the effect of the experimental treatment is the same at randomization in the test arm as at drop-in in the control arm. In addition, their model assumes that patients who receive drop-in therapy are comparable to those who do not, although the authors noted that baseline covariates could be incorporated, and thus their model could include factors that predict drop-in. Shao et al. (2005) extended the Branson & Whitehead (2002) methodology to allow the effect of drop-in to vary with time receiving the drop-in therapy, and also defined a latent hazard rate model with the same features. White (2006) noted that the recensoring procedure of Branson & Whitehead (2002) needs to be modified when the control arm survival time without drop-in depends on the drop-in time, otherwise a bias towards the null results if drop-in patients have a poor prognosis, and away from the null if they have a good prognosis.

White (2006) also pointed out that the estimation procedure proposed in Shao et al. (2005) is biased when the drop-in time is prognostic. For the structural nested modelling approaches, such as those of Branson & Whitehead (2002) and Shao et al. (2005), one major concern is that the assumed model for the true survival time does not account for the disparity that some subjects experience the intermediate event while others do not. Furthermore, assuming a constant experimental treatment effect between treatment arms may be questionable since the disease course is more advanced among drop-in patients that receive delayed therapy.

In this paper, we tackle this practical problem from a completely different modelling perspective than the aforementioned methods. Instead of modelling the true survival time using either the accelerated failure time model or the proportional hazards model, we model the observed event times using a semiparametric hazards model. To account for the fact that some subjects do not experience the intermediate event, we introduce a mixture model to characterize the progression and nonprogression subpopulation. Furthermore, for the progression population, we separately model the time to the intermediate event and the time from the intermediate event to death. We also include baseline covariates and prognostic covariates in both time-to-event regression models. In this way, we not only account for the heterogeneity at baseline, but also capture the heterogeneity at treatment switching. Finally, our model assumes a parametric switching effect at the time of the intermediate event, which may be different from the baseline treatment effect. The advantages of our model are clear: we model only observed event times which makes it possible to assess model assumptions and check model fit using the observed data; we allow the treatment effect at switching to be completely different from the baseline treatment effect; and the model can handle both baseline covariates and prognostic covariates at switching.

2. Panitumumab study

Our proposed methodology was motivated by the panitumumab colorectal cancer clinical trial conducted by Amgen Inc. (Amado et al., 2008). This clinical trial was an open label, randomized, phase III multicentre study designed to compare the efficacy and safety of panitumumab plus best supportive care versus best supportive care alone in colorectal cancer patients. One objective was to compare the treatment effect on the overall survival time in this subject population.

Subjects were randomly assigned to receive treatment or control. Panitumumab was administered until disease progression, inability to tolerate the investigational product, or other reasons for discontinuation. During the study, subjects in the control group who had disease progression at any time were eligible to receive panitumumab at 6 mg/kg administered once every 2 weeks as part of a separate protocol. Figure 1 shows the counts for each group in the follow-up period. Among the 223 patients on the control arm, 201 patients had disease progression, of which 167 switched over to the treatment arm. Due to this substantial switching percentage, this study provides strong motivation for developing new statistical models as well as new methods for estimating the true causal effect of the treatment in the presence of a semicompeting risk.

3. Proposed method

3·1. *Models and assumptions*

In cancer clinical trials, some subjects experience the intermediate event of disease progression and others do not and these subjects are censored for the event. To address this issue, we propose a mixed semicompeting risks transition model. We assume that the population consists of two subpopulations, where one population will eventually develop disease progression before death, but the other population will never experience disease progression. For the no-progression

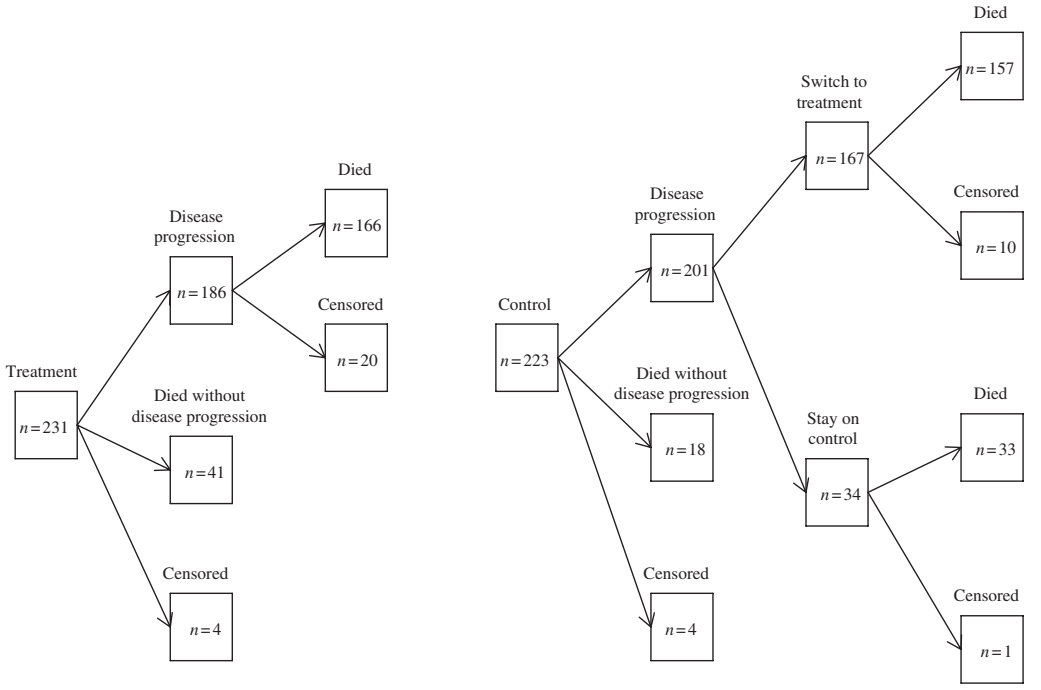

Fig. 1. Graphical representation of the panitumumab data: total $n = 454$.

population, the only event time of interest is time-to-death, but for the progression population, both the time to disease progression and the event of death need to be considered.

To introduce our statistical models, we use the following notation: a dichotomous variable U is used to denote the lifetime disease progression status of the subjects, $U = 1$ if the subject has disease progression before death and 0 otherwise; we let T_{D} denote the time to death for the no-progression subjects with $U = 0$; for the other subjects with $U = 1$, we use T_U to denote their time to disease progression and let G denote the time from disease progression to death.

The proposed statistical model has three components. The first component models the distribution of the progression status given the baseline covariates X and randomized treatment R:

$$\operatorname{logit}\{\operatorname{pr}(U = 1 \mid R, X)\} = \alpha_0 + \alpha_1 R + \alpha_2 X, \tag{1}$$

where $R = 1$ if the patient is on the experimental treatment arm and 0 otherwise, and the αs are unknown regression coefficients. The second component models the survival distribution for the no-progression population given X and R:

$$h_{\mathrm{D}}(t \mid R, X, U = 0) = h_0(t) \exp(\beta_0 R + \gamma_0 X), \tag{2}$$

where $h_{\mathrm{D}}(t \mid R, X, U = 0)$ is the conditional hazard function of T_{D} given the covariates, $h_0(t)$ is an unknown baseline hazard function and $(\beta_0, \gamma_0^{\mathrm{T}})^{\mathrm{T}}$ are unknown regression coefficients. In the third component, we model the distributions of time to disease progression, T_U, and time from disease progression to death, G, in the progression population, given treatment switching or not, by assuming a transition model structure:

$$\begin{aligned} h_U(t \mid R, X, U = 1) &= h_1(t) \exp(\beta_1 R + \gamma_1 X), \\ h_G(t \mid R, Z, V, U = 1, T_U) &= h_2(t) \exp\{\beta_{21} R + \beta_{22} V(1 - R) + \gamma_2^{\mathrm{T}}(Z^{\mathrm{T}}, T_U)^{\mathrm{T}}\}, \end{aligned} \tag{3}$$

where $h_U(t \mid R, X, U=1)$ is the conditional hazard function for T_U, $h_G(t \mid R, Z, U=1, T_U)$ is the conditional hazard function of G, both $h_1(t)$ and $h_2(t)$ are unknown baseline hazard functions, and βs and γs are regression coefficients. Here, V indicates the treatment switching and Z, which contains X, reflects covariates collected at baseline and at disease progression, which could be prognostic factors for the switching decision.

Our models naturally account for the situation that some subjects may or may not experience disease progression for reasons other than ignorable censoring and also include the gap time between disease progression and death, thereby automatically implying that disease progression occurs before death. Because we condition on the disease progression status, our model can be considered as a type of pattern-mixture model. As one reviewer points out, our method can also be viewed as an illness-death model with four states: alive with/without progression, and dead with/without progression (Fix & Neyman, 1951; Sverdrup, 1965), but with a pattern-mixture parameterization (Larson & Dinse, 1985). However, our current hazard models have different interpretations in the survival context. For example, all the models are the hazard models for some specific survival events so that all the β parameters give the treatment effects on the risk of these events. Specifically, β_{21} represents the logarithm of the hazard ratio of treatment post-disease progression while the coefficient of T_U gives the effect of the disease progression time on future death. In the second model of (3), since the switching only happens to some subjects on the control arm; $V(1-R)$ is used in the regression. Furthermore, our model (3) assumes that, for the same subjects, the hazard function after the switching at disease progression would change by $\exp(\beta_{22})$, when compared with the case when they had no switching. Obviously, the latter is a structural assumption, which is acceptable in practice.

Our goal is to compare the survival function of death time in the setting when no subjects have switching. To see how to use the proposed models for achieving this goal, we adopt a counterfactual outcome framework by defining $T_{\mathrm{D}}^*(a)$ as a potential survival time when a subject receives treatment a and never changes treatment status and letting $S_a(t) = \mathrm{pr}\{T_{\mathrm{D}}^*(a) > t\}$. Thus, we are interested in comparing $S_1(t)$ and $S_0(t)$. As in the usual causal framework, we assume the following consistency assumption and no unobserved confounder assumption:

Assumption 1. Treatment R is completely randomized and $T_{\mathrm{D}}^*(a) = T_{\mathrm{D}}(a)$ if a subject never changes treatment.

Assumption 2. Given $(R=0, Z, T_U=s)$, that is, a subject in the control arm has disease progression at time s and covariates Z, or $(R=1, Z, T_U=s)$, V is independent of the potential outcomes $\{T_{\mathrm{D}}^*(0), T_{\mathrm{D}}^*(1)\}$.

Let $f_X(x)$ and $f_Z(z)$ denote the density functions for X and Z, respectively. Then by the randomization of R, we obtain the potential survival function of treatment a, $\mathrm{pr}\{T_{\mathrm{D}}^*(a) > t\}$, as

$$
\begin{aligned}
&\mathrm{pr}\{T_{\mathrm{D}}^*(a) > t \mid R=a\} \\
&\quad = \int_x \mathrm{pr}\{T_{\mathrm{D}}^*(a) > t \mid X, U=0, R=a\}\mathrm{pr}(U=0 \mid X, R=a) f_X(x \mid R=a)dx \\
&\qquad + \int_{x,z,s} \mathrm{pr}\{T_{\mathrm{D}}^*(a) > t \mid T_U=s, Z, U=1, R=a\}d\mathrm{pr}(T_U \leqslant s \mid Z, U=1, R=a) \\
&\qquad \times f_Z(z \mid X, U=1, R=a)\mathrm{pr}(U=1 \mid X, R=a) f_X(x \mid R=a)dzdx.
\end{aligned}
$$

From Assumption 2, $\mathrm{pr}\{T_{\mathrm{D}}^*(a) > t \mid T_U=s, X, Z, U=1, R=a\} = \mathrm{pr}\{T_{\mathrm{D}}^*(a) > t \mid V=0, T_U=s, X, Z, U=1, R=a\}$. On the other hand, $(R=a, U=0)$ or $(R=a, U=1, V=0)$ implies

that the treatment status is never switched so $T_{\mathrm{D}}^{*}(a)$ can be replaced by the observed T_{D} in the above expression by Assumption 1. Since $T_{\mathrm{D}} = G + T_U$ for subjects with $U = 1$, we obtain the survival functions $S_a(t)$ as follows:

$$
\begin{aligned}
&\int_x \mathrm{pr}(T_{\mathrm{D}} > t \mid X, U = 0, R = a)\mathrm{pr}(U = 0 \mid X, R = a) f_X(x \mid R = a)dx \\
&\quad + \int_{x,z,s} \mathrm{pr}(G > t - s \mid T_U = s, V = 0, Z, U = 1, R = a)d\mathrm{pr}(T_U \leqslant s \mid Z, U = 1, R = a) \\
&\quad \times f_Z(z \mid X, U = 1, R = a)\mathrm{pr}(U = 1 \mid X, R = a) f_X(x \mid R = a)dzdx \\
&= \int_x \mathrm{pr}(T_D > t \mid X, U = 0, R = a)\mathrm{pr}(U = 0 \mid X, R = a) f_X(x \mid R = a)dx \\
&\quad + \int_{x,z} \Bigg\{ \mathrm{pr}(T_U > t \mid Z, U = 1, R = 1) \\
&\quad + \int_0^t \mathrm{pr}(G > t - s \mid T_U = s, V = 0, Z, U = 1, R = a)d\mathrm{pr}(T_U \leqslant s \mid Z, U = 1, R = a) \Bigg\} \\
&\quad \times f_Z(z \mid X, U = 1, R = a)\mathrm{pr}(U = 1 \mid X, R = a) f_X(x \mid R = a)dzdx.
\end{aligned}
$$

In other words, $S_a(t)$ can be expressed in terms of the parameters in our models (1)–(3) and the distributions of X and Z given $(X, U = 1, R)$. Hence, by inserting the estimates of these parameters into the above expression, we will be able to estimate $S_a(t)$, and thus the causal effect of treatment.

In real applications, there is often some potential bias due to censoring and obtaining the differential prognostic covariates at disease progression. To eliminate such bias, we need the following assumptions:

Assumption 3. The censoring time is independent of T_{D}, G and T_U given the observed covariates.

Assumption 4. For progression subjects, T_U is independent of Z given R and X.

Assumptions 3 and 4 discard the contribution of the censoring distribution. Assumption 4 is plausible if the part of Z excluding X is collected after disease progression.

3·2. *Inference procedure*

Let Y denote the observed event if no disease progression occurs; otherwise, we use Y to denote the second event time and W to denote the disease progression time. Let Δ be the censoring indicator. The observed data can be divided into four groups of observations:

Group 1. Subjects are observed to die at time Y and no disease progression has been observed. Clearly, these subjects belong to the first subpopulation with $U = 0$ and $T_{\mathrm{D}} = Y$. The observed data are $(Y, \Delta = 1, U = 0, X, R)$. Thus, the contribution to the likelihood function is $h_0(t)\exp(\beta_0 R + \gamma_0 X)\exp\{-H_0(t)e^{\beta_0 R + \gamma_0 X}\}\mathrm{pr}(U = 0 \mid R, X) f_X(x \mid R)\mathrm{pr}(R)$.

Group 2. Subjects are observed to have disease progression at W and die at Y. These subjects belong to the second subpopulation $(U = 1)$ and $T_U = W$, $T_{\mathrm{D}} = Y$ so $G = Y - W$. The observed data are $(T_U, G, \Delta = 1, U = 1, V, Z, X, R)$. Thus, the contribution to the

likelihood function is

$$h_1(W)\exp(\beta_1 R+\gamma_1 X)\exp\{-H_1(W)\exp(\beta_1 R+\gamma_1 X)\}h_2(G)\exp\{\beta_{21}R+\beta_{22}V(1-R)+\gamma_2(Z,W)\}\exp[-H_2(G)\exp\{\beta_{21}R+\beta_{22}V(1-R)+\gamma_2(Z,W)\}]f_Z(z\mid X,R,U=1)\times\mathrm{pr}(U=1\mid R,X)f_X(x\mid R)\mathrm{pr}(R),$$

where $\mathrm{pr}(Z\mid X,R,U=1)=\mathrm{pr}(Z \text{ without } X\mid X,R,U=1)$. We will use this notation for all conditional distributions of Z given X thereafter.

Group 3. Subjects are observed to have disease progression at W and censored at C. The subjects belong to the second subpopulation with $U=1$ and $T_U=W$, $T_{\mathrm{D}}>C=Y$ so $G>Y-W$. The observed data are $(T_U, G>Y, \Delta=0, U=1, V, Z, X, R)$. Thus, the contribution to the likelihood function is

$$h_1(W)\exp(\beta_1 R+\gamma_1 X)\exp\{-H_1(W)\exp(\beta_1 R+\gamma_1 X)\}\exp[-H_2(Y-W)\exp\{\beta_{21}R+\beta_{22}V(1-R)+\gamma_2(Z,W)\}]f_Z(z\mid X,R,U=1)\mathrm{pr}(U=1\mid R,X)f_X(x\mid R)\mathrm{pr}(R).$$

Group 4. Subjects are only observed to be censored at Y and no disease progression occurs before Y. These subjects may belong to the first subpopulation, $U=0$, with $T_{\mathrm{D}}>Y$; or, they may belong to the second subpopulation, $U=1$, with $T_U>Y$. The observed data are $\{UT_U+(1-U)T_{\mathrm{D}}>Y, X, R\}$. Thus, the contribution to the likelihood function is $[\exp\{-H_0(Y)e^{\beta_0 R+\gamma_0 X}\}\mathrm{pr}(U=0\mid R,X)+\exp\{-H_1(Y)e^{\beta_1 R+\gamma_1 X}\}\mathrm{pr}(U=1\mid R,X)]f_X(x\mid R)\mathrm{pr}(R)$.

For inference, we estimate all the model parameters, including the βs, γs and Hs, via the nonparametric maximum likelihood approach. In this approach, the baseline hazard functions, H_0, H_1 and H_2, are assumed to be step functions with jumps at the observed event times. To compute the nonparametric maximum likelihood estimates, we will use the expectation-maximization algorithm to facilitate the computation of the nonparametric maximum likelihood estimates. Specifically, we treat U_i for subject i as potential missing data. Then it is clear that only for subjects in Group 4, U_i is not observed. To estimate the asymptotic covariance matrix of the parameter estimates, we treat all the αs, βs, γs and the jump sizes of the Hs as parameters and use their observed information matrix. In particular, the observed information matrix can be calculated using the Louis formula (Louis, 1982) and its inverse is used as the estimator for the asymptotic covariance matrix.

3·3. *Prediction of the survival function with partial treatment crossover*

To estimate $S_a(t)$, we can estimate each term on the right-hand side of $S_a(t)$, given in § 3·1, using the parameter estimates. Specifically, the estimators are

$$\hat{\mathrm{pr}}(T_{\mathrm{D}}>t\mid R=a,X,U=0)=\exp\left\{-\hat{H}_0(t)\exp(\hat{\beta}_0 R+\hat{\gamma}_0 X)\right\},$$

$$\hat{\mathrm{pr}}(U=0\mid R=a,X)=\left\{1+\exp(\hat{\alpha}_0+\hat{\alpha}_1 R+\hat{\alpha}_2 X)\right\}^{-1},$$

$$\hat{f}_X(x\mid R=a)=\sum_{j=1}^{n} I(X_j=x,R_j=a)/\sum_{j=1}^{n} I(R_j=a),$$

$$\hat{\mathrm{pr}}(T_U>t\mid Z,U=1,R=1)=\exp\left\{-\hat{H}_1(t)\exp(\hat{\beta}_1 a+\hat{\gamma}_1 X)\right\},$$

$$\hat{\mathrm{pr}}(G>s'\mid T_U=s,V=0,Z,U=1,R=a)=\exp\left[-\hat{H}_2(s')\exp\{\hat{\beta}_{21}a+\hat{\gamma}_2(Z,s)\}\right],$$

$$\hat{f}_Z(z \mid X, R=a, U=1) = \frac{\sum_{j\in(\text{groups 2, 3})} \delta(Z_j = z) K_{a_n}(\|X_j - X\|) I(R_j = a)}{\sum_{j\in(\text{groups 2, 3})} K_{a_n}(\|X_j - X\|) I(R_j = a)},$$

where in the final expression, δ is the Dirac delta function and K_{a_n} is a kernel weight with bandwidth a_n. Either the delta method or resampling techniques like the bootstrap will be used to derive the confidence band of $\hat{S}_a(t)$. As demonstrated in the simulation studies in § 5, we find, in practice, that the bootstrap is easy to implement and performs well for moderate sample sizes. In addition, our experience also shows that even though kernel estimation of $\text{pr}(Z \mid X, R, U=1)$ may be biased if the dimension of the Xs is not small, the final estimate of $S_a(t)$ is not that sensitive to the dimension of X due to the averaging operations used in calculating $\hat{S}_a(t)$.

To compare the experimental arm survival function with the control arm survival function without switching, we examine the weighted difference $n\sum_{k=1}^{K} \hat{\omega}(t_k)\{\hat{S}_1(t_k) - \hat{S}_0(t_k)\}^2$, where $t_1, \ldots, t_K$ are prespecified time-points in $[0, \tau]$ and $\hat{\omega}(t_k)$ is a weight specified at time t_k. Useful weight functions $\hat{\omega}(t)$ can be based on the class of $\hat{S}_0(t)^{\rho_1}\{1 - \hat{S}_0(t)\}^{\rho_2}$, where ρ_1 and ρ_2 are constants in $[0, 1]$ and $\hat{S}_0(t)$ can be also replaced by $\hat{S}_1(t)$ or a pooled estimator of survival functions. Thus, by choosing different ρ_1 and ρ_2, we can emphasize comparisons at either early stages or late stages of follow-up. In the subsequent analysis, we consider $(\rho_1, \rho_2) = (0, 1)$ or $(1/2, 1/2)$. Under the null hypothesis for which $S_0(t) = S_1(t)$, according to the asymptotic results to be given later, $\sqrt{n}\{\hat{S}_1(t_1) - \hat{S}_0(t_1), \ldots, \hat{S}_1(t_K) - \hat{S}_0(t_K)\}^{\mathrm{T}} \to N(0, \Sigma)$ in distribution for some covariance matrix Σ. Thus, it is easy to see that $n\sum_{k=1}^{K} \hat{\omega}(t_k)\{\hat{S}_1(t_k) - \hat{S}_0(t_k)\}^2$ converges in distribution to $\mathcal{Z}^{\mathrm{T}}\Sigma^{1/2}\text{diag}\{\omega(t_1), \ldots, \omega(t_K)\}\Sigma^{1/2}\mathcal{Z}$, where $\mathcal{Z}$ denotes a multivariate standard normal variate and $\omega(t)$ is the limit of $\hat{\omega}(t)$. We reject the null hypothesis if the test statistic is larger than the $(1-\alpha)$-percentile of $\mathcal{Z}^{\mathrm{T}}\hat{\Sigma}^{1/2}\text{diag}\{\hat{\omega}(t_1), \ldots, \hat{\omega}(t_K)\}\hat{\Sigma}^{1/2}\mathcal{Z}$, where $\hat{\Sigma}$ is a consistent estimator for Σ.

Remark 1. Because of the randomization, $\text{pr}(X \mid R=a) = \text{pr}(X)$, and therefore, we can replace $\hat{\text{pr}}(X \mid R=a)$ with the empirical distribution of X. However, we observe very little efficiency gain in numerical studies.

4. ASYMPTOTIC PROPERTIES

We establish the asymptotic properties for the parameter estimators and $\hat{S}_a(t)$ using the general nonparametric maximum likelihood theory framed in Zeng & Lin (2010). In addition to Assumptions (1)–(4), we need the following assumptions:

Assumption 5. The true parameters values of the βs, γs and αs, still denoted as $\theta \equiv (\beta_0, \beta_1, \beta_{21}, \beta_{22}, \gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}}, \alpha_0, \alpha_1, \alpha_2^{\mathrm{T}})^{\mathrm{T}}$, belong to a bounded set in real Euclidean space. Moreover, the true baseline functions, still denoted as (h_0, h_1, h_2), are continuous and are bounded away from zero in $[0, \tau]$, where τ is the study duration.

Assumption 6. If there is some constant ν such that $\nu^{\mathrm{T}}(1, R, Z) = 0$ with probability one, then $\nu = 0$. Additionally, we assume (R, Z) to have bounded support and there exits a continuous component of X such that its coefficient in model (1) is nonzero.

Assumption 7. With probability one, $\text{pr}(C \geqslant \tau \mid R, Z) > 0$ and $\text{pr}(V = 1 \mid R = 0, Z, T_U) \in (\mu_0, \mu_1)$ for some constant $0 < \mu_0 < \mu_1 < 1$.

Under these conditions, the following theorems give the consistency and asymptotic distribution of the estimators.

THEOREM 1. *Under Assumptions* (1)–(7), $\|\hat{\theta} - \theta\| + \sum_{k=1}^{3} \sup_{t\in[0,\tau]} | \hat{H}_k(t) - H_k(t) | \to 0$, *almost surely, where* $\hat{\theta} \equiv (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\gamma}_0, \hat{\gamma}_1^{\mathrm{T}}, \hat{\gamma}_2^{\mathrm{T}}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2^{\mathrm{T}})^{\mathrm{T}}$ *and* $\hat{H}_k(t)$ *is the estimator of* $H_k(t)$.

THEOREM 2. *Under Assumptions* (1)–(7), $\sqrt{n}(\hat{\theta} - \theta, \hat{H}_1 - H_1, \hat{H}_2 - H_2, \hat{H}_3 - H_3)$ *converges in distribution to a mean zero Gaussian process in the metric space* $R^d \times l^\infty[0,\tau] \times l^\infty[0,\tau] \times l^\infty[0,\tau]$, *where* d *is the dimension of* θ.

Using the results from Theorem 1 and Theorem 2, we can further obtain the asymptotic distribution of $\hat{S}_a(t)$ as given in the previous section.

THEOREM 3. *In addition to Assumptions* (1)–(7), *we assume that* $K_{a_n}(x) = a_n^{-d_x} K(\|x\|/a_n)$, *where* d_x *is the dimension of* X, $K(\cdot)$ *is a symmetric kernel density function with* $\int y^s K(y)dy = 0, s = 1, \ldots, (m-1)$ *with* $m > d/2$, *and* a_n *satisfies* $na_n^d \to \infty$, $na_n^{2m} \to 0$. *Then with probability one,* $\sup_{t\in[0,\tau]} | \hat{S}_a(t) - S_a(t) | \to 0$ *and for each fixed* t, *and* $\sqrt{n}\{\hat{S}_a(t) - S_a(t)\}$ *converges in distribution to a mean zero normal variate.*

5. SIMULATION STUDIES

5·1. *Simulation study I*

To examine the small sample performance of the proposed method, we conducted a simulation study by generating data from models (1)–(3). Specifically, the baseline treatment $R = 1$ for the first half of the subjects and 0 for the other half; two baseline covariates X_1 and X_2 are independently generated from the uniform distribution on $[-1, 1]$, and a Bernoulli with success probability 0·6, respectively. We then use models (2) and (3) to further generate time to events of interest. The susceptibility status is Bernoulli with success probability $1/\{1+\exp(-1{\cdot}6+1{\cdot}8R - X_1 - 0{\cdot}1X_2)\}$. For the no-progression subjects with $U = 0$, we simulate their death time T_{D} using model (2) with $H_0(t) = t$, $\beta_0 = -1$ and $(\gamma_{01}, \gamma_{02}) = (1, 0{\cdot}2)$. For the progression subjects, the time to disease progression, T_U, is generated from the first hazard model in (3) with $H_1(t) = t/2$, $\beta_1 = -0{\cdot}5$ and $(\gamma_{11}, \gamma_{12}) = (1, 0)$. Finally, to generate the time from disease progression to death for the progression subjects with $U = 1$, we first generate the prognostic factors Z at disease progression from the uniform distribution on [0, 1]. The assignment to treatment switching, V, in the untreated subjects is assumed to have a Bernoulli with success probability $1/\{1+\exp(0{\cdot}5 - 0{\cdot}3T_U - 0{\cdot}2X_1 - 0{\cdot}5Z\}$, yielding a switching rate of 38·7% in the control arm. Then, the time from disease progression to death, G, follows the second hazard model in (3) with $H_2(t) = \exp(t) - 1$, $\beta_{21} = -0{\cdot}3$, $\beta_{22} = -0{\cdot}5$, and $\gamma_{21} = 0{\cdot}6$, $\gamma_{22} = -0{\cdot}5$, $\gamma_{23} = 0{\cdot}5$, $\gamma_{24} = -0{\cdot}4$. Thus, the subjects who change treatment status from untreated to treated will have their hazard risk reduced by $\exp(0{\cdot}5)$ and the longer the time disease progression is, the longer the survival time will be. Finally, the censoring time is generated from a uniform distribution on (1, 7) and the study duration is $\tau = 3$. The latter yields average proportions for groups 1 to 4 as 23, 41, 21 and 13%.

In the simulation study, we consider sample sizes of $n = 400$ and $n = 1000$. Bootstrap samples of size 50 are used to construct pointwise 95% confidence intervals for the estimated survival probability. The results from 1000 replicates are given in Table 1. Additionally, we calculate the square root of the mean square error and maximum absolute difference of the estimated survival curve using 200 equally spaced time-points between 0 and the maximum censoring time. In particular, for the methods of intent-to-treat, Branson & Whitehead (2002), Shao et al. (2005), and the proposed model, the square roots of the mean square errors of $\hat{S}_0(t)$ are 0·062, 0·050, 0·054

Table 1. *Simulation study I*

		$n = 400$				$n = 1000$			
Parameter	True	EST	SD	ESE	CP%	EST	SD	ESE	CP%
				Survival model of no-progression population					
β_0	−1·0	−1·08	0·26	0·26	94	−1·04	0·16	0·16	93
γ_{01}	1·0	0·96	0·23	0·22	94	0·97	0·13	0·13	94
γ_{02}	0·2	0·22	0·23	0·24	94	0·20	0·14	0·14	95
				Disease progression model of progression population					
β_1	−0·5	−0·54	0·16	0·16	94	−0·52	0·10	0·10	94
γ_{11}	1·0	1·08	0·14	0·14	91	1·04	0·09	0·08	94
γ_{12}	0·0	−0·01	0·14	0·15	94	0·00	0·09	0·09	95
				Gap time model of progression population					
β_{21}	−0·3	−0·31	0·19	0·19	95	−0·30	0·12	0·12	95
β_{22}	−0·5	−0·50	0·20	0·20	95	−0·51	0·12	0·12	96
γ_{21}	0·6	0·61	0·18	0·18	95	0·60	0·11	0·11	95
γ_{22}	−0·5	−0·51	0·16	0·16	94	−0·50	0·10	0·10	95
γ_{23}	0·5	0·51	0·17	0·17	94	0·50	0·10	0·10	95
γ_{24}	−0·4	−0·41	0·27	0·28	95	−0·41	0·17	0·17	95
				Susceptibility model					
α_0	1·6	1·66	0·26	0·26	95	1·63	0·16	0·16	96
α_1	−1·8	−1·80	0·27	0·28	94	−1·79	0·17	0·16	96
α_{21}	1·0	0·90	0·24	0·24	93	0·94	0·15	0·15	95
α_{22}	0·1	0·11	0·26	0·27	94	0·11	0·16	0·15	95
				Predicted survival functions in control arm					
$S_0(\tau/2)$	0·51	0·49	0·04	0·04	91	0·49	0·02	0·02	92
$S_0(\tau)$	0·17	0·18	0·03	0·03	91	0·18	0·02	0·02	91
				Predicted survival functions in experimental arm					
$S_1(\tau/2)$	0·63	0·61	0·03	0·03	94	0·61	0·02	0·02	92
$S_1(\tau)$	0·32	0·33	0·03	0·03	93	0·33	0·02	0·02	94

EST, average of the parameter estimates; SD, sample standard deviation of the estimates; ESE, average of the standard error estimates; CP%, coverage probability of the 95% confidence interval based on a normal approximation.

and 0·035, respectively; the square root of the mean square errors of $\hat{S}_1(t)$ are 0·032, 0·028, 0·030 and 0·030, respectively; the maximum absolute differences of $\hat{S}_0(t)$ are 0·101, 0·067, 0·083 and 0·051, respectively and those of $\hat{S}_1(t)$ are 0·061, 0·036, 0·051 and 0·049, respectively.

In addition, we study the power of the test statistics proposed in § 3·3 under the simulation setup discussed above, and obtain the Type I error rate by letting $\beta_0 = \beta_1 = \beta_{21} = \beta_{22} = \alpha_1 = 0$. In particular, we consider two test statistics, test I with $(\rho_1, \rho_2) = (0, 1)$ and test II with $(\rho_1, \rho_2) = (0{\cdot}5, 0{\cdot}5)$, for sample sizes of $n = 400$ and $n = 1000$. When $n = 400$, the powers are 87·1% for test I, and 85·6% for test II. The power increases to 99·8% for both tests when the sample size increases to 1000. For tests I and II, the Type I error rates are 4·5% and 4·8% when $n = 400$, respectively; and 5·4% and 4·8% when $n = 1000$, respectively.

5·2. *Simulation study II*

In the second simulation study, we use the modified simulation setup as in Shao et al. (2005) to compare the proposed model with existing methods for switching, and to also demonstrate the robustness of the proposed model. In particular, the survival time is generated according to the exponential distribution with hazard rate 0·0693 for the control arm and 0·0462 for the experimental treatment arm. The total sample size of $n = 600$, with 300 subjects in each arm. For both treatment arms, the random censoring time is generated according to the uniform

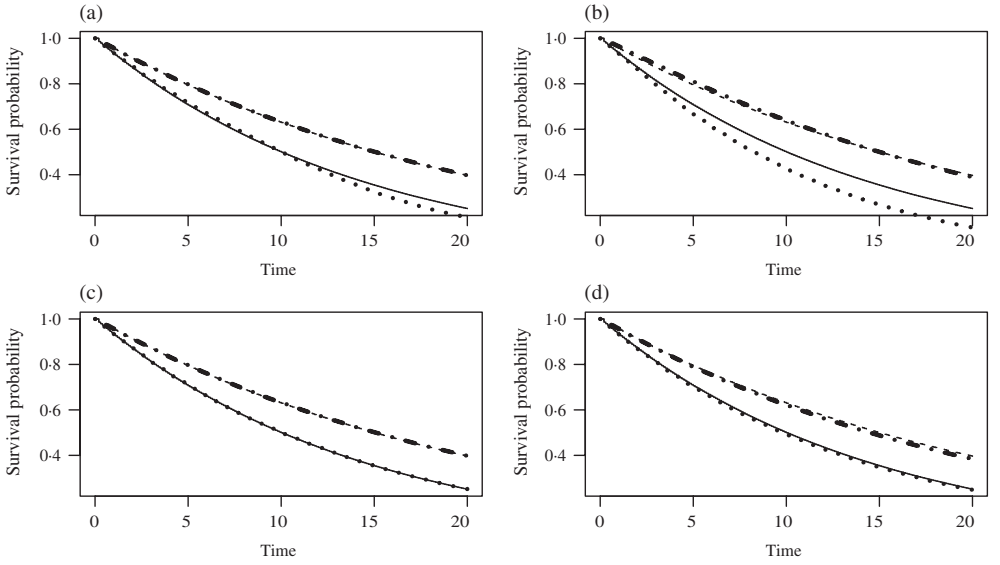

Fig. 2. Average predicted survival curves from simulation study II. $S(t \mid R = a)$ is the potential survival function for subjects with treatment status a if they have no treatment switching. In each panel, the solid curve is the true survival function in the control arm, the dashed curve is the true survival function in the experimental treatment arm, while the dotted and the dash-dotted curves are, respectively, the estimated survival functions in these two arms. The estimates in the plots are based on (a) intent-to-treat analysis, (b) Branson & Whitehead (2002), (c) Shao et al. (2005) and (d) our method.

distribution on the interval of 15–20 months, resulting in an overall censoring percentage of 35·4%. The time-to-event, which is the switching time in Shao et al. (2005), is generated from the exponential distribution with a mean of 7·22 months for the control group and 10·82 months for the experimental treatment group. In this paper, we focus on switching only from the control to treatment arm and let the patients switch at the event time with probability 0·6, yielding a switching rate of 39·0%. After switching, the observed survival time for the switching patients is updated using equation (4) and (11) in Shao et al. (2005) with $\beta = -0{\cdot}4055$, $\eta_{00} = 0{\cdot}1$, $\eta_{01} = 0{\cdot}009$ and $\eta_{10} = \eta_{11} = 0$.

Figure 2 presents the averaged predicted survival curves on $[0, 20]$ for the intent-to-treat Kaplan–Meier, Branson & Whitehead (2002), Shao et al. (2005) and proposed methods using 1000 simulations. The discrepancy in the four different approaches focuses on the estimation of the control group survival curve. In particular, the survival curves generated by Shao et al. (2005) and the proposed model yield survival curves with smaller bias. The square roots of the mean square errors and maximum absolute differences are 0·029 and 0·049, respectively, for Shao et al. (2005), and 0·031 and 0·043, respectively, for the proposed approach. On the other hand, bigger biases are observed for the Branson & Whitehead (2002) and the intent-to-treat approaches. The square roots of the mean square errors and maximum absolute differences are, respectively, 0·073 and 0·088 for the Branson & Whitehead (2002) method, and 0·034 and 0·069 for the intent-to-treat approach.

6. Analysis of the panitumumab data

We carry out here a detailed analysis of the panitumumab study. It is purely an abstract construct of the semicompeting risk nature of the proposed model to assume the existence of a subpopulation that is subject to disease progression, and thus this condition is not assumed to

apply literally to this study. The baseline covariates we consider are initial treatment, age in years at screening, baseline electrocorticography performance status with 0 or 1 versus $\geqslant 2$, primary tumour diagnosis type with rectal versus colon, gender, and region with three levels consisting of western Europe, eastern and central Europe, and rest of the world. In the panitumumab study, the median age was 62·5 years and the interquartile range of age was (55, 69) years. There were 388 patients with electrocorticography score 0 or 1, 287 were male, 151 had rectal cancer, 352 were from Western Europe, 39 were from Eastern and Central Europe, and 63 were from the rest of the world. The median follow-up time was 189·5 days and the interquartile range of the follow-up time was (93, 334) days. Among those 387 patients who developed disease progression, the median disease progression time is 53 days and the interquartile range is (45, 84) days.

The model for the time of disease progression includes all the baseline covariates. Among the 387 patients who developed disease progression, the median age at the time of disease progression was 62·1 years with interquartile range (55·0, 69·1), the numbers of patients who had partial response, stable disease and progressive disease were 19, 86 and 282, respectively. There were 348 patients with baseline electrocorticography score 0 or 1, 286 patients had a last electrocorticography score 0 or 1, and 180 patients had grade 2 or above adverse events.

The covariates at the time from disease progression to death include additional progonostic factors for the switching decision. They are progression time, partial response age, best tumour response with partial response or stable disease versus progressive disease according to investigator assessment, last electrocorticography performance status and grade 2 or above adverse events. We include those prognostic factors based on our best knowledge with assistance from trial clinicians so that Assumption 2 could be valid. Because of the dependency on the unobserved outcome, Assumption 2 is not testable, and the results could be biased if it is violated.

The results from the proposed model are given in Table 2. The survival probability estimates using the intent-to-treat Kaplan–Meier, no switching subgroup analysis using Kaplan–Meier, Branson & Whitehead (2002), Shao et al. (2005), and the proposed methods at the 25, 50, 75 and 100% quartiles of time to death are given in Table 3. The no-switching approach excludes the patients who switched from best supportive care alone to panitumumab plus supportive care. The p-values to test the treatment effect using the above five approaches are 0·577, $<$0·001, 0·520, 0·002 and $<$0·001, respectively. For the Branson & Whitehead (2002) method, 1000 bootstrap samples are used to construct the standard error and p-value because the standard errors calculated from the covariance matrix at convergence are too small to construct valid confidence limits (Branson & Whitehead, 2002). Figure 3 provides the predicted survival curves for the two treatment groups of panitumumab plus best supportive care and best supportive care alone using the five approaches. We notice that the intent-to-treat Kaplan–Meier and Branson & Whitehead (2002) approaches yield small survival differences between the panitumumab plus best supportive care and best supportive care alone groups before 200 days and there is little difference after 200 days since enrolment. On the other hand, the subgroup analysis based on no patients switching shows big differences between the two arms for the whole follow-up period. We then investigated the reason for the survival curve discrepancy for best supportive care alone group and found two key contributing factors: a high switching rate for the best supportive care alone group, 167/223=75%; and selection bias, that is, the patients with longer time-to-event were more likely to be switched from best supportive care alone to panitumumab plus best supportive care with a median time-to-event of 40·5 days for the no switching patients and 49 days for the switching patients. The Shao et al. (2005) and the proposed approach both yield big differences in the estimated survival curves compared with the control arm with the bigger treatment difference being obtained by the proposed approach.

Table 2. *Model estimates for the panitumumab data*

Parameter	EST	SE	*P*	Parameter	EST	SE	*P*
	T_D Model				T_U Model		
Treatment	−0·464	3·47	0·182	Treatment	−1·144	1·18	<0·001
Age	0·023	0·15	0·124	Age	−0·015	0·05	0·004
bECOG	−0·589	2·99	0·048	bECOG	−0·805	1·74	<0·001
Rectal	−0·028	3·20	0·929	Rectal	−0·018	1·10	0·871
Male	−0·288	3·05	0·345	Male	−0·054	1·09	0·622
CenEastEU	−0·188	6·27	0·764	CenEastEU	0·194	2·50	0·439
WesternEU	0·181	3·99	0·650	WesternEU	−0·068	1·60	0·672
	T_G Model				U Model		
Treatment	−0·784	2·14	<0·001	Intercept	1·366	9·72	0·160
V*(1-Treatment)	−1·383	2·09	<0·001	Treatment	−1·070	3·19	<0·001
Prog Time	−0·003	0·01	0·039	Age	−0·008	0·14	0·546
PR Age	−0·004	0·05	0·450	bECOG	1·905	3·34	<0·001
BTR PR	−0·226	3·45	0·512	Rectal	0·314	3·31	0·342
BTR SD	−0·180	1·74	0·302	Male	−0·303	3·21	0·346
bECOG	−0·268	1·96	0·173	CenEastEU	0·078	6·23	0·901
LECOG	−1·035	1·48	<0·001	WesternEU	0·346	4·12	0·400
AE	0·295	1·16	0·011				

EST, parameter estimates; SD(×10), standard error of the estimates; *P*, *p*-values; bECOG, baseline electrocorticography; CenEastEU, central Europe; WesternEU, western Europe; Prog Time, progression time; PR, partial response; BTR, best tumour response; SD, stable disease; LECOG, last electrocorticography; AE, adverse event.

Table 3. *Predicted survival functions for the panitumumab data*

Time	ITT		No Crossover		IPE		Shao Cox		TM	
(Days)	BSC	P+BSC	BSC	P+BSC	BSC	P+BSC	BSC	P+BSC	BSC	P+BSC
93	0·750	0·793	0·303	0·793	0·722	0·783	0·678	0·798	0·548	0·801
190	0·511	0·533	0·081	0·533	0·454	0·551	0·341	0·536	0·171	0·555
334	0·266	0·260	0·020	0·260	0·201	0·298	0·097	0·258	0·025	0·282
1024	0·013	0·038	0·020	0·038	0·001	0·007	0·001	0·023	0·001	0·026
p-value	0·577		<0·001		0·520		0·002		<0·001	

ITT, intent-to-treat; IPE, Branson & Whitehead (2002); Shao Cox, Shao et al. (2005); TM, proposed method; BSC, best supportive care alone; P+BSC, panitumumab plus best supportive care.

7. Extensions

We have conducted simulation studies when data are generated from the proposed model or from the models in Shao et al. (2005). Additional simulation studies may be carried out to further examine the robustness of the proposed method to misspecification of models (1)–(3) or to perform sensitivity analysis on the ignorability assumption of switching selection.

Although the proposed model is developed under partial treatment switching, i.e., not all patients switch treatment, it can be easily extended to the case with complete treatment switching in which all patients on the control arm switch to the experimental arm. Specifically, under complete treatment switching, the components of the proposed model for U, T_D and T_U remain the same and only the model in (3) for G needs to be modified as follows

$$h_G(t \mid R, Z, V, U = 1, T_U) = h_2(t) \exp\{\beta_{21} R + \gamma_2(Z, T_U)\},$$

as in this case β_{22} is no longer identifiable. However, under complete treatment switching, the estimator of the predictive survival function needs to be rederived, which is much more

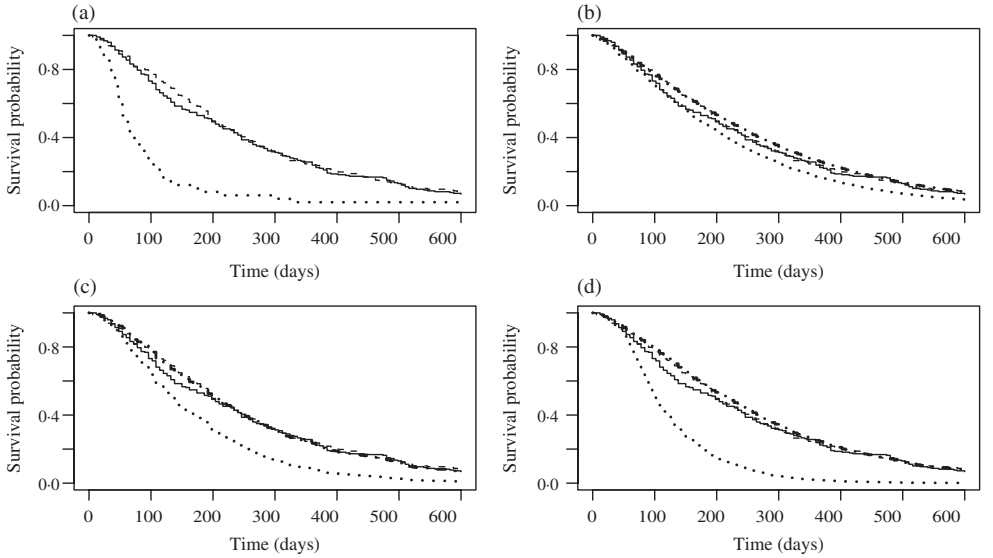

Fig. 3. Predicted survival curves for the panitumumab data. In each panel, the solid curve is the intent-to-treat survival function in the control arm, the dashed curve is the intent-to-treat survival function in the experimental treatment arm, while the dotted and the dash-dotted curves are, respectively, the estimated survival functions in these two arms. The estimates in the survival curves are based on (a) no switching subgroup analysis using the Kaplan–Meier estimates, (b) the Branson & Whitehead (2002) method, (c) the Shao et al. (2005) method and (d) our method.

challenging than under partial treatment switching. In addition, in the proposed model, we build dependence between T_U and G via the transition model. An alternative to the transition model is the frailty model. Under the latter, we have $h_U(t \mid R, X, U = 1) = h_1(t) \exp(\beta_1 R + \gamma_1 X)\omega$, and $h_G(t \mid R, Z, V, U = 1, T_U) = h_2(t) \exp\{\beta_{21} R + \beta_{22} V(1 - R) + \gamma_2 Z\}\omega$, where ω is a latent gamma-frailty with mean one and variance θ. Compared with the frailty model, the transition model is much more numerically stable in the implementation of the expectation-maximization algorithm. Finally, the proposed method can also be extended to the case where patients may switch from either treatment arm as discussed in Shao et al. (2005). These extensions, along with comparison between the transition model and the frailty model, are currently under investigation.

ACKNOWLEDGEMENT

This research work was part of a collaborative effort with Amgen, Inc. and was partly funded by the National Institutes of Health, U.S.A. and by Amgen Inc. We thank the editor, associate editor and the two referees for their constructive comments and suggestions which have greatly improved this paper. The Amgen Research Group consists of, in alphabetical order, Drs Chao-Yin Chen, Eric M. Chi, Thomas Liu, Jean Pan, Steve M. Snapinn, Mike Wolf, Allen Xue and Nan Zhang. The authors do not have a conflict of interest in any portions of this paper.

APPENDIX

Proof of Theorem 1. Let $l_n(\theta, H_1, H_2, H_3)$ denote the observed loglikelihood function for (θ, H), and $H\{t\} = \{H(t) - H(t-)\}$. First, it is easy to see if $\hat{H}_k\{t\} = \infty$, then $l_n(\hat{\theta}, \hat{H}_1, \hat{H}_2, \hat{H}_3) = -\infty$. Moreover, this also holds if the jump size of $\hat{H}_k$ at the corresponding events is zero. Thus, the jump sizes of $\hat{H}_k$ at the corresponding events are positive and finite so the derivatives of $l_n(\theta, H_1, H_2, H_3)$ with respect to each

jump size of $\hat{H}_k$ should be zero at $(\hat{\theta}, \hat{H}_1, \hat{H}_2, \hat{H}_3)$. This gives

$$\hat{H}_0\{Y_i\} = I_i(G_1)/\left\{\sum_{j\in G_1} e^{\hat{\beta}_0 R_j + \hat{\gamma}_0^{\mathrm{T}} X_j} + \sum_{j\in G_4} \frac{(1-\hat{p}_{uj})\hat{S}_{\mathrm{D}}(Y_i)}{(1-\hat{p}_{uj})\hat{S}_{\mathrm{D}}(Y_i) + \hat{p}_{uj}\hat{S}_U(Y_i)}\right\}, \tag{A1}$$

$$\hat{H}_1\{W_i\} = I_i(G_2, G_3)/\left\{\sum_{j\in G_2, G_3} e^{\hat{\beta}_1 R_j + \hat{\gamma}_1^{\mathrm{T}} X_j} + \sum_{j\in G_4} \frac{\hat{p}_{uj}\hat{S}_U(Y_i)}{(1-\hat{p}_{uj})\hat{S}_{\mathrm{D}}(Y_i) + \hat{p}_{uj}\hat{S}_U(Y_i)}\right\}, \tag{A2}$$

$$\hat{H}_2\{W_i\} = I_i(G_2)/\left\{\sum_{j\in G_2, G_3} e^{\hat{\eta}_{Gj}}\right\}, \tag{A3}$$

where G_i denotes group i, $I_i(A) = I(i \in A)$, $\hat{\eta}_{\mathrm{D}} = \hat{\beta}_0 R + \hat{\gamma}_0^{\mathrm{T}} X$, $\hat{\eta}_U = \hat{\beta}_1 R + \hat{\gamma}_1^{\mathrm{T}} X$, $\hat{\eta}_G = \hat{\beta}_{21} R + \hat{\beta}_{22} V(1-R) + \hat{\gamma}_2^{\mathrm{T}}(Z, W)$, and the additional subindex j denotes the expression for the jth subject. In addition, we let $\hat{S}_{\mathrm{D}}(t) = \exp\{-\hat{H}_0(t)e^{\hat{\eta}_{Dj}}\}$, $\hat{S}_U(t) = \exp\{-\hat{H}_1(Y_i)e^{\hat{\eta}_{Uj}}\}$ and $\hat{p}_{uj} = \hat{\mathrm{pr}}(U = 1 \mid R_j, X_j)$. Equation (A3) implies $\hat{H}_0\{Y_i\} \leqslant I_i(G_1)/\sum_{j\in G_1} c_0$, where c_0 is a positive lower bound of $e^{\hat{\eta}_{Dj}}$. Since $n^{-1}\sum_{j\in G_1} 1 \to \mathrm{pr}(U = 0, Y \leqslant C) > 0$, we obtain $\limsup_n \hat{H}_0(\tau) \leqslant \limsup_n n^{-1}\sum_{i=1}^n I(Y_i \leqslant \tau, i \in G_1)/c_0 n^{-1}\sum_{j\in G_1} 1 < \infty$. Similarly, equations (A2) and (A3) yield that $\limsup_n \hat{H}_1(\tau)$ and $\limsup_n \hat{H}_2(\tau)$ are both finite.

By Helly's selection theorem, for any subsequence, we can choose a further subsequence such that $\hat{H}_k$ weakly converges to an increasing function H_k^* for $k = 1, 2, 3$. Moreover, we can assume $\hat{\theta} \to \theta^*$. We then show $H_k^* = H_k$ and $\theta^* = \theta$. To this end, we construct $\tilde{H}_k$ such that $\tilde{H}_k$ has jumps at the same events as $\hat{H}_k$; moreover, the jumps of $\tilde{H}_k$ are given by the right-hand side of (A1) to (A3) except that the parameters on the right-hand side are set to be the true values. It is straightforward to verify that $\tilde{H}_k$ converges uniformly to the true function H_k. Furthermore, we can show that $d\hat{H}_k/d\tilde{H}_k$ converges uniformly to dH_k^*/dH_k.

Therefore, since $l_n(\hat{\theta}, \hat{H}_1, \hat{H}_2, \hat{H}_3) - l_n(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_3) \geqslant 0$, we take limits on both sides and conclude that the Kullback–Leilber information between $(\theta^*, H_1^*, H_2^*, H_3^*)$ and (θ, H_1, H_2, H_3) is nonpositive. This immediately implies that the loglikelihood function at $(\theta^*, H_1^*, H_2^*, H_3^*)$ is equal to the loglikelihood function at (θ, H_1, H_2, H_3) with probability one. Thus, this equality holds for all subjects in Groups 1 to 4 as defined in § 3. Comparing the differences of the loglikelihood functions from subjects in Group 2 and Group 3, we have

$$(H_2^*)'(G)e^{\beta_{21}^* R + \beta_{22}^* V(1-R)\gamma_2^{*\mathrm{T}}(Z,W)} = H_2'(G)e^{\beta_2 R + \beta_{22} V(1-R)\gamma_2^{\mathrm{T}}(Z,W)},$$

so by Assumption 6, $H_2^* = H_2$, $\beta_{21}^* = \beta_{21}$, $\beta_{22}^* = \beta_{22}$ and $\gamma_2^* = \gamma_2$. Let $W = 0$, we have

$$\{h_1^*(0)e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}\}/(1 + e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}) = \{h_1(0)e^{\alpha_0 + \alpha_1 R + \alpha_2{}^{\mathrm{T}} X}\}/(1 + e^{\alpha_0 + \alpha_1 + \alpha_2{}^{\mathrm{T}} X}).$$

Now in the loglikelihood for subjects in Group 1, we let $Y = 0$ and obtain

$$h_0^*(0)/(1 + e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}) = h_0(0)/(1 + e^{\alpha_0 + \alpha_1 + \alpha_2{}^{\mathrm{T}} X}).$$

Compare the above equations, so $\alpha_1^* = \alpha_1, \alpha_2^* = \alpha_2$. Since one component of X is continuous and has a nonzero coefficient in α_2, the above equation gives $h_0^*(0) = h_0(0)$ and $\alpha_0^* = \alpha_0$. Finally, after integrating the likelihood equality function for Group 2 for W from 0 to Y, we have

$$\frac{[1 - \exp\{-H_1^*(Y)e^{\beta_1^* R + \gamma_1^{*\mathrm{T}} X}\}]e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}}{1 + e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}} = \frac{[1 - \exp\{-H_1(Y)e^{\beta_0 R + \gamma_1{}^{\mathrm{T}} X}\}]e^{\alpha_0^* + \alpha_1^* R + \alpha_2^{*\mathrm{T}} X}}{1 + e^{\alpha_0 + \alpha_1 + \alpha_2{}^{\mathrm{T}} X}}.$$

Thus, $H_1^* = H_1$ and $\beta_1^* = \beta_1, \gamma_1^* = \gamma_1$. On the other hand, integrating the likehood equality function for subjects in Group 1 for Y from 0 to Y gives

$$\frac{1-\exp\{-H_0^*(Y)e^{\beta_0^* R+\gamma_0^{*\mathrm{T}}X}\}}{1+e^{\alpha_0^*+\alpha_1^* R+\alpha_2^{*\mathrm{T}}X}} = \frac{1-\exp\{-H_0(Y)e^{\beta_0 R+\gamma_0{}^{\mathrm{T}}X}\}}{1+e^{\alpha_0+\alpha_1+\alpha_2{}^{\mathrm{T}}X}},$$

so $\beta_0^* = \beta_0$, $\gamma_0^* = \gamma_0$ and $H_0^* = H_0$.

We have proved that $\hat{\theta} \to \theta$ and $\hat{H}_k$ converges weakly to H_k. The latter can be further strengthened to uniform convergence in $[0, \tau]$ since H_k is continuous. Therefore, Theorem 1 holds. □

Proof of Theorem 2. The proof of Theorem 2 follows from the same argument in proving Theorem 2 in Zeng & Lin (2010). In particular, their conditions (C.1)–(C.4) and (C.6) hold for our specific models. Their first identifiability condition (C.5) has been verified in the proof of Theorem 1. To complete the proof, it remains to verify the second identifiability of their condition (C.7). Consider the score function along a sub model $H_k + \epsilon \int f_k dH_k$ and $\theta + \epsilon\nu$ where $\nu = (\beta_0, \gamma_0, \beta_1, \gamma_1, \beta_{21}, \beta_{22}, \gamma_2, \alpha_0, \alpha_1, \alpha_2)$. If this score function is zero with probability one, then we need to show that $f_k = 0$ and $\nu = 0$. For subjects in Group 2, the score equation is

$$\begin{aligned} 0 = {} & f_1(W) + \eta_U - \int_0^W f_1(t)dH_1(t)e^{\eta_U} - H_1(Y)e^{\eta_U}\eta_U + f_2(G) + \eta_G \\ & - \int_0^G f_2(t)dH_2(t)e^{\eta_G} - H_2(Y)e^{\eta_G}\eta_G + e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X}(1+e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X})^{-2} \\ & \times (\xi_0 + \xi_1 R + \xi_2^{\mathrm{T}}X). \end{aligned} \tag{A4}$$

For subjects in Group 3, we obtain the score equation to be

$$\begin{aligned} 0 = {} & f_1(W) + \eta_U - \int_0^W f_1(t)dH_1(t)e^{\eta_U} - H_1(Y)e^{\eta_U}\eta_U - \int_0^G f_2(t)dH_2(t)e^{\eta_G} \\ & - H_2(Y)e^{\eta_G}\eta_G + e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X}(1+e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X})^{-2}(\alpha_0 + \alpha_1 R + \alpha_2^{\mathrm{T}}X). \end{aligned} \tag{A5}$$

The difference between (A4) and (A5) gives $f_2(G) + \eta_G = 0$, so by Assumption 6, $f_2 = 0$, $\beta_{21} = 0$, $\beta_{22} = 0$ and $\gamma_2 = 0$.

Using this result and equation (A5), the score equation for subjects in Group 4 becomes

$$0 = -\int_0^Y f_0(t)dH_0(t)e^{\eta_{\mathrm{D}}} - H_0(Y)e^{\eta_{\mathrm{D}}}\eta_{\mathrm{D}} - \frac{e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X}(\alpha_0 + \alpha_1 R + \alpha_2^{\mathrm{T}}X)}{(1+e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X})^2}. \tag{A6}$$

On the other hand, for subjects in Group 1,

$$0 = f_0(Y) + \eta_{\mathrm{D}} - \int_0^Y f_0(t)dH_0(t)e^{\eta_{\mathrm{D}}} - H_0(Y)e^{\eta_{\mathrm{D}}}\eta_{\mathrm{D}} - \frac{e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X}(\alpha_0 + \alpha_1 R + \alpha_2^{\mathrm{T}}X)}{(1+e^{\alpha_0+\alpha_1 R+\alpha_2^{\mathrm{T}}X})^2}. \tag{A7}$$

Then the difference between (A6) and (A7) gives $f_0(Y) + \eta_{\mathrm{D}} = 0$ which further gives $f_0 = 0$, $\beta_0 = 0$ and $\gamma_0 = 0$. As a result, (A7) becomes $\alpha_0 + \alpha_1 R + \alpha_2^{\mathrm{T}}X = 0$ so $\alpha_0 = 0$, $\alpha_1 = 0$ and $\alpha_2 = 0$. This further combined with equation (A5) gives $f_1 = 0$, $\beta_1 = 0$ and $\gamma_1 = 0$. We have verified condition (C.7) in Zeng & Lin (2010). According to their results, our Theorem 2 holds.

Moreover, from Theorem 3 in Zeng & Lin (2010), we also conclude that the inverse of the observed information is a consistent estimator for the asymptotic covariance. □

Proof of Theorem 3. The consistency of $\hat{S}_a(t)$ follows from the consistency of the following terms, $\hat{\mathrm{pr}}(T_{\mathrm{D}} > t \mid R = a, X, U = 0)$ $\hat{\mathrm{pr}}(U = 0 \mid R = a, X)$, $\hat{\mathrm{pr}}(G + T_U > t \mid R = a, X, Z, U = 1)$, from

Theorem 1. Moreover, we have the fact that, by the kernel approximation,

$$\frac{\sum_{j\in G_2,G_3}\hat{\text{pr}}(G+T_U>t\mid R=a,X_j,Z_j,U=1)K_{a_n}(X_j-x)I(R_j=a)}{\sum_{j\in G_2,G_3}K_{a_n}(X_j-x)I(R_j=a)}$$
$$\to E\{\text{pr}(G+T_U>t\mid R=a,X,Z,U=1)\mid X=x,R=a,U=1,T_U\leqslant C\}$$

uniformly in x in the support of X and with probability one. Since Z and (T_U,C) are independent given (R,X), the limit on the right-hand side is also equal to $\text{pr}(G+T_U>t\mid X=x,R=a,U=1)$. Thus, $\hat{S}_a(t)\to S_a(t)$. Note $\hat{S}_a(t)-S_a(t)$ can be written as

$$\begin{aligned}
&\frac{1}{n_a}\sum_{i=1}^{n}\hat{\text{pr}}(T_{\text{D}}>t\mid R=a,X=X_i,U=0)\hat{\text{pr}}(U=0\mid R=a,X_i)I(R_i=a)\\
&\quad+\frac{1}{n_a}\sum_{i=1}^{n}E\left\{\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)\Big|X=X_i,R=a,U=1,T_U<C\right\}\\
&\quad\times\hat{\text{pr}}(U=1\mid X_i,R=a)I(R_i=a)-S_a(t)\\
&\quad+\frac{1}{n_a}\sum_{i=1}^{n}\left[\frac{\sum_{j\in G_2,G_3}\hat{\text{pr}}(G+T_U>t\mid R=a,X_j,Z_j,U=1)K_{a_n}(X_j-X_i)I(R_j=a)}{\sum_{j\in G_2,G_3}K_{a_n}(X_j-X_i)I(R_j=a)}\right.\\
&\quad-\left.E\left\{\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)\Big|X=X_i,R=a,U=1,T_U<C\right\}\right]\\
&\quad\times\hat{\text{pr}}(U=1\mid X_i,R=a)I(R_i=a).
\end{aligned}\tag{A8}$$

The first two terms are Hadamard differentiable with respect to $\hat{\theta}$, $\hat{H}_k$ and the empirical distribution of X given R. Therefore, by the functional delta method, these terms can be approximated as $\Sigma(\hat{\theta}-\theta)+\sum_{k=1}^{3}\int f_k(t)d\{\hat{H}_k(t)-H_k(t)\}+\int g(x)d\{\hat{F}(X\mid R)-F(X\mid R)\}+o_p(n^{-1/2})$ for some bounded functions $f_k(t)$ and $g(x)$, where $\hat{F}(X\mid R)$ is the empirical distribution function of X given R. To complete the proof of Theorem 3, we only need to show that the last term in equation (A8) is asymptotically normal.

Denote Q as subjects in Group 2 and Group 3 and use P_n to denote the empirical measure. The last term of (A8) can be reorganized as

$$\begin{aligned}
&(P_n-P)\left[\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)I(R=a)\right.\\
&\qquad\left.\times\,Qn_a^{-1}\left\{\sum_{i=1}^{n}\frac{K_{a_n}(X-X_i)}{n^{-1}\sum_{k=1}^{n}Q_kK_{a_n}(X_k-X_i)I(R_k=a)}\right\}\right]\\
&-(P_n-P)\left[n_a^{-1}\sum_{i=1}^{n}\frac{E\left\{\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)I(R=a)QK(X-X_i)\right\}}{E\{QK_{a_n}(X-X_i)I(R=a)\}^2}\right.\\
&\qquad\left.\times\,\hat{\text{pr}}(U=1\mid X_i,R=a)I(R_i=a)\right]\\
&+n_a^{-1}\sum_{i=1}^{n}\left[\frac{E\left\{\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)I(R=a)QK(X-X_i)\right\}}{E\{QK_{a_n}(X-X_i)I(R=a)\}}\right.\\
&\quad\left.-\,U\left\{\hat{\text{pr}}(G+T_U>t\mid R=a,X,Z,U=1)\Big|X=X_i,R=a,U=1,T_U<C\right\}\right]\\
&\qquad\times\hat{\text{pr}}(U=1\mid X_i,R=a)I(R_i=a).
\end{aligned}\tag{A9}$$

In (A9), we can apply the functional central limit result in Theorem 2.11.23 of van der Vaart & Wellner (1996) to show that the first two terms of (A9) converge in distribution to a Gaussian process with a factor $n^{1/2}$. From the kernel approximation, the last term is $O(a_n^m)$, and therefore, is $o_p(n^{-1/2})$. Combining the above results, we conclude that Theorem 3 holds. □

REFERENCES

AMADO, R. G., WOLF, M., PEETERS, M., CUTSEM, E. V., SIENA, S., FREEMAN, D. J., JUAN, T., SIKORSKI, R., SUGGS, S., RADINSKY, R., et al. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **28**, 1626–34.

BARTHEL, F. M. S., BABIKER, A., ROYSTON, P. & PARMAR, M. K. B. (2006). Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statist. Med.* **25**, 2521–42.

BRANSON, M. & WHITEHEAD, J. (2002). Estimating a treatment effect in survival studies in which patients switch treatment. *Statist. Med.* **21**, 2449–63.

FIX, E. & NEYMAN, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Hum. Biol.* **23**, 205–41.

GREENLAND, S., LANES, S. & JARA, M. (2008). Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and g-estimation. *Clin. Trials* **5**, 5–13.

HERNÁN, M. Á., BRUMBACK, B. & ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11**, 561–70.

JIANG, Q., SNAPINN, S. M. & IGLEWICZ, B. (2004). Calculation of sample size in endpoint trials: the impact of informative noncompliance. *Biometrics* **60**, 800–6.

LACHIN, J. M. & FOULKES, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* **42**, 507–19.

LAKATOS, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **44**, 229–41.

LARSON, G. & DINSE, G. (1985). A mixture model for the regression analysis of competing risks data. *Appl. Statist.* **34**, 201–11.

LAW, M. G. & KALDOR, J. M. (1996). Survival analyses of randomized trials adjusting for patients who switch treatments. *Statist. Med.* **15**, 2069–76.

LONDON, W. B., FRANTZ, C. N., CAMPBELL, L. A., SEEGER, R. C., BRUMBACK, B. A., COHN, S. L., MATTHAY, K. K., CASTLEBERRY, R. P. & DILLER, L. (2010). Phase II randomized comparison of topotecan plus cyclophosphamide versus topotecan alone in children with recurrent or refractory neuroblastoma: a children's oncology group study. *J. Clin. Oncol.* **28**, 3808–15.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **13**, 2233–47.

LU, J. & PAJAK, T. F. (2000). Statistical power for a long-term survival trial with a time-dependent treatment effect. *Contr. Clin. Trials* **21**, 561–73.

LUNCEFORD, J. K., DAVIDIAN, M. & TSIATIS, A. A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48–57.

MARCUS, S. M. & GIBBONS, R. D. (2001). Estimating the efficacy of receiving treatment in randomized clinical trials with noncompliance. *Health Serv. Outcomes Res. Methodol.* **2**, 247–58.

PORCHER, R., LÉVY, V. & CHEVRET, S. (2002). Sample size correction for treatment crossovers in randomized clinical trials with a survival endpoint. *Contr. Clin. Trials* **23**, 650–61.

ROBINS, J. M. & TSIATIS, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun. Statist.* A **20**, 2609–31.

ROBINS, J. M., HERNÁN, M. Á. & BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–60.

SHAO, J., CHANG, M. & CHOW, S.-C. (2005). Statistical inference for cancer trials with treatment switching. *Statist. Med.* **24**, 1783–90.

SVERDRUP, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfersb etween differents tates of health. *Skand. Aktuar.* **52**, 185–211.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

WHITE, I. R. (1997). Editorial on survival analyses of randomized trials adjusting for patients who switch treatments by M. G. Law and J. M. Kaldor. *Statist. Med.* **16**, 2619–25.

WHITE, I. R. (2006). Letter to the editor. Estimating treatment effects in randomized trials with treatment switching. *Statist. Med.* **25**, 1619–22.

WHITE, I. R., CARPENTER, J., POCOCK, S. J. & HENDERSON, R. A. (2003). Adjusting treatment comparisons to account for non-randomized interventions: an example from an angina trial. *Statist. Med.* **22**, 781–93.

YAMAGUCHI, T. & OHASHI, Y. (2004). Adjusting for differential proportions of second-line treatment in cancer clinicals trials. Part II: an application in a clinical trial of unresectable non-small-cell lung cancer. *Statist. Med.* **23**, 2005–22.

ZENG, D. & LIN, D. Y. (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statist. Sinica* **20**, 871–910.

[*Received November* 2010. *Revised August* 2011]