J. R. Statist. Soc. B (2007)
69, *Part* 4, *pp.* 565–588

Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination

Jack Cuzick, Peter Sasieni, Jonathan Myles and Jonathan Tyrer

Wolfson Institute of Preventive Medicine, London, UK

[Received February 2005. Final revision February 2007]

Summary. Methods for adjusting for non-compliance and contamination, which respect the randomization, are extended from binary outcomes to time-to-event analyses by using a proportional hazards model. A simple non-iterative method is developed when there are no covariates, which is a generalization of the Mantel–Haenszel estimator. More generally, a 'partial likelihood' is developed which accommodates covariates under the assumption that they are independent of compliance. A key feature is that the proportion of contaminators and non-compliers in the risk set is updated at each failure time. When covariates are not independent of compliance, a full likelihood is developed and explored, but this leads to a complex estimator. Estimating equations and information matrices are derived for these estimators and they are evaluated by simulation studies.

Keywords: Non-compliance; Partial likelihood; Proportional hazards model; Randomized clinical trials; Semiparametric models

1. Introduction

It is conventional to analyse randomized trials according to the treatment option that is assigned to each patient—the so-called intent-to-treat method. The dangers of excluding non-compliant patients or analysing a trial according to actual treatment are well recognized (Altman (1991), pages 461–471) and can lead to biases which undermine the original reasons for randomization. The main problem is that compliers often have a different underlying hazard failure from that of non-compliers. This has been clearly established in both screening (Duffy *et al.*, 2002) and treatment (Bonadonna and Valagussa, 1981) trials. However, the intent-to-treat approach dilutes the effects of treatment in compliers by the uninformative outcomes that are seen in non-compliers. It can thus lead to a substantial underestimate of the value of a treatment when applied to individual patients who are willing to accept it. This can be particularly relevant in situations where consent is sought after randomization, such as large scale population screening trials.

By extending an approach that was used by Sommer and Zeger (1991), we previously developed a method for evaluating the effect of a new treatment on binary outcomes, which adjusted for refusal to accept the new treatment (non-compliance) and/or off-protocol use of the treatment among controls (contamination). This approach fully respects the randomization but allows an unbiased estimate of the value of treatment in those patients who are willing to accept it (Cuzick *et al.*, 1997). With this method the magnitude of the treatment effect is usually larger

Address for correspondence: Jack Cuzick, Cancer Research UK Centre for Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, Queen Mary College, London, EC1M 6BQ, UK.
E-mail: jack.cuzick@cancer.org.uk

 1369–7412/07/69565

than that found by intent-to-treat methods, but the confidence intervals are also wider, so the asymptotic power levels remain the same. An additional advantage is that more realistic confidence intervals are provided, which is particularly important for trials that are aimed at showing the equivalence of two treatments.

The previous work focused on binary outcomes, and the object of the present paper is to extend this approach to the case in which the time to event is of interest. In Cuzick *et al.* (1997) we indicated a somewhat *ad hoc* approach based on stratification of time intervals. Robins and Tsiatis (1991) and others have developed methods that are based on accelerated failure models. However, the most widely used method for analysing clinical trials with time to event as an end point is based on the proportional hazards model (Cox, 1972), and it is desirable to extend this standard method to deal with non-compliance and contamination. Some work in this direction has been carried out by Loeys and Goetghebeur (2003), who studied the case of non-compliance only (without contamination) in the absence of covariates. They did not assume proportionality between the base-line hazards for the two groups but could not compute asymptotic properties of their estimator for the treatment effect.

We focus primarily on the case where the decision to opt for the non-allocated treatment occurs at the time of randomization. In view of the symmetry between non-compliance and contamination, we shall not further distinguish between these two. Individual patient covariates are easily modelled, and this type of analysis is particularly appropriate for the Zelen (1979, 1990) type of randomization model, where consent is sought after randomization. By redefining 'compliance' to mean 'response to treatment', the model can also be used to analyse the survival benefits in patients who respond to treatment (or receive a full course of treatment) in an unbiased way. This is currently a question which is dealt with in an unsatisfactory and biased way, as responders to treatment are likely to be biased towards having better survival, even in the absence of an effect of treatment (Peto *et al.*, 1976; Altman, 1991).

We do not consider the more complicated situation in which patients might wish to use a third treatment, although these methods could be extended to this situation. Here we focus on comparisons of a new treatment with a standard treatment (which may be no treatment), and non-compliance is defined as opting for the standard treatment when allocated the new treatment. The method also requires some modification for placebo-controlled trials. If we assume that the decision not to comply is made early on (and independent of side-effects), then non-compliers in the two arms can be treated equally. This is discussed briefly in the final section.

Another area where these methods are of interest is in disease screening, where prolonged follow-up is necessary to determine the effect of screening on disease prevalence or mortality. In Duffy *et al.* (2002) our method for binary outcomes has been used to estimate the effect of non-compliance in breast screening trials, and to use this to adjust case–control analyses of on-going screening programmes to allow for this effect. On the basis of available data from five trials, this analysis suggested that, on average, non-compliers to an invitation for breast screening have a 36% higher breast cancer mortality rate than uninvited controls.

The paper proceeds by first developing the model and notation. A likelihood is developed and then specialized to the case in which any base-line survival differences in compliers or contaminators can be modelled as being proportional adjustments to the base-line hazard. This likelihood proves difficult to solve analytically, and some simpler cases are considered first. When there are no covariates a non-iterative estimate is developed. This estimate contains arbitrary weights and the optimal, as well as some simple, weighting systems are studied. Next a 'partial likelihood' is introduced. Its estimating equations and variance estimators are deduced and more easily

solved than that for the full likelihood. A further section evaluates these methods numerically, first in a simple example and then via simulations. The last section discusses limitations and extensions. A full list of all notation is given in Table 1 for easy reference.

2. Basic model and notation

As before (Cuzick *et al.*, 1997), it is convenient to think of the trial population as composed of three classes of individuals (Fig. 1):

(a) *insistors*—subjects who demand the new treatment regardless of the group to which they are randomized,
(b) *ambivalent*—subjects who accept whichever treatment they are offered and
(c) *refusers*—subjects who refuse the new treatment, if offered.

Below we shall use the letters I, A and R to refer to these groups in equations, etc. Each of these classes is split at random to receive the new treatment (T) or the conventional treatment (C). The statistical difficulties arise because it is not always possible to determine which (latent) class individuals belong to. Thus, only four groups can be directly observed from the data (Fig. 1):

(i) (denoted CT)—insistors randomized to control;
(ii) (denoted CC)—ambivalent and refusers randomized to control;
(iii) (denoted TT)—insistors and ambivalents randomized to new treatment;
(iv) (denoted TC)—refusers randomized to new treatment.

Let $\lambda_k(t)$ and $\Lambda_k(t)$ be the hazard and cumulative hazard functions in the absence of treatment for $k = \mathrm{A}, \mathrm{R}$ and in the presence of treatment for $k = \mathrm{I}$ (insistors). We shall assume that these are proportional, i.e. $\lambda_k(t) = \exp(\gamma_k)\,\lambda(t)$, with $\gamma_\mathrm{A} = 0$ (i.e. $\lambda \equiv \lambda_\mathrm{A}$) for identifiability. Let $f_k(t)$ and $S_k(t)$ be the corresponding density and survivor functions. Let $z_0 \equiv z_{0i}$ be the indicator for the ith subject to be randomized to new treatment, and to be in class A, and let $z \equiv (z_1, \ldots, z_m)$ be any other covariates. Finally let δ_i be the indicator for subject i being an observed failure time (0 corresponding to a censored observation) and assume that censoring is independent of and non-informative for the failure time process. In particular, censoring is independent of allocated treatment for each of the classes of individuals. The basic model then postulates that an individual with covariates (k, z_0, z) will have hazard

$$\exp(\gamma_\mathrm{T} z_0 + \beta z + \gamma_k)\,\lambda(t). \tag{1}$$

Note that the treatment effect γ_T is only apparent for patients in class A and that z_0, and in some cases k, are not observable covariates. All insistors receive treatment regardless of the

	Randomize	
	Control	**T**reatment
Contaminators (**I**nsistors)	CT	TT
Ambivalents	CC	TT
Non-compliers (**R**efusers)	CC	TC

Fig. 1. Groups referred to in the analysis classified by randomization and class of individual

Table 1. Notation

Notation	*Definition*
a_{ji}^k	$\pi_k S_{ji}^k / \{\pi_k S_{ji}^k + (1-\pi_k) S_{ji}^{\mathrm{A}}\}$
A_{ki}^l	Terms in the expansion of the covariance matrix for the partial likelihood
d_i	Normalizing constant
D_i^σ	Number of events (failures) in group σ at time t_i
$f_k(t)$	Density function
$F_k(z)$	(Base-line) distribution of z in latent class k
I_{ji}	$I_{\{T_j \geqslant t_i\}}$
K_i	Normalizing constants (for each failure time)
$\hat{K}_i$	Empirical estimator of K_i
L_i	Likelihood contribution for individual i (given $i \in \sigma$)
n_{ki}^l	Number of individuals in class k randomized to l who are at risk at t_i-
N_i^σ	Number at risk in group σ at time t_i-
p_i^k	$\tau_{ki}^{\delta_i} S_i^k$
R_i	Set of individuals at risk at t_i-
S_i^σ	$\Sigma_{j \in \sigma} S_{ji}$
S_{ji}^σ	$E(I_{ji} \mid l \in \sigma)$
S_{ji}^k	Survival probability to t_i given latent class k and covariate z_j
$S_k(t)$	Survival function
$\{t_i\}$	Ordered failure times (no ties)
$\{T_j\}$	Ordered observation times (1 per individual)
$v = (v_c) = (v_{lc})$	3×6 matrix of 0s and 1s used to express the full likelihood scores
v_{ki}	v_{kii}
$\overline{v_{ki}}$	Weighted average of v_{kji} over individuals in R_i with weights θ_{ji}
v_{kji}	Derivative of θ_{ji} with respect to γ_k: probability that individual $j \in R_i$ has hazard ratio $\exp(\gamma_k)$ given j is in group σ
$\overline{v}_{li}$	$\Sigma_{j \geqslant i,c}\, v_{lc}\theta_{ljc} / \Sigma_{j \geqslant i,c}\, \theta_{ljc}$
v_{li}^*	$\Sigma_c\, v_{lc} w_{ic}$
w_i, w_{ki}	$\Im_{t_i-}$-measurable weights
w_{ic}	Term in the full likelihood score equivalent to the probability that an individual i (with covariate z_i and in R_i and σ) contributes to a given component (c) of the vector score.
z	Covariates
$\overline{z_i}$	Weighted average of z_j ($j \in R_i$) with weights θ_{ji}
z_0	Indicator of randomized to treatment and ambivalent
β	Regression parameters for z
γ	$(\gamma_{\mathrm{T}}, \gamma_{\mathrm{I}}, \gamma_{\mathrm{R}})$
γ_k	Log-hazard-ratio
δ_j	Indicator of individual j not being censored
Δ_i	Jump in (estimated) base-line cumulative hazard at t_i
Δ_{kl}	Kronecker's delta
θ_i	θ_{ii}
θ_{ji}	(Partial likelihood) hazard ratio for individual j at time t_i, dependent on β and γ
θ_k	Hazard ratio
θ_{ljc}	$w_{jc} \exp(\gamma v_c + \beta z_j)$
$\lambda(t)$	$\lambda_{\mathrm{A}}(t)$
$\lambda_k(t)$	Hazard function
$\Lambda_k(t)$	Cumulative hazard function
μ_k	Parameter defined by $\mathrm{d}F_k(z) = \exp(\mu_k z)\, \mathrm{d}F_{\mathrm{A}}(z)$
$\pi_{\mathrm{I}i}$	Estimate of the proportion of insistors among those at risk at t_i- who had been randomized to treatment and complied
$\overline{\pi}_{\mathrm{I}i}$	Conditional probability of being an insistor given in both TT and R_i
π_k	π_{ki} evaluated at time 0
$\pi_k(z)$	Probability of the event in π_k conditional also on covariate z

(*continued*)

Table 1 (*continued*)

Notation	*Definition*
π_{Ri}	Estimate of the proportion of refusers among those at risk at t_i- who had been randomized to control and complied
$\overline{\pi}_{Ri}$	Conditional probability of being a refuser given in both CC and R_i
ρ	Randomization ratio (treated:control)
τ_{kj}	Hazard ratio given latent class k and covariate z_j
$\Im_{t_i-}$	σ-field generated by all events before t_i the failure(s) at time t_i
Subscript $k =$ I, A, R (T)	I, insistor; A, ambivalent; R, refuser; T, treatment
Superscript $l =$ T, C	Refers to randomization (T, treatment; C, control)
Superscript $\sigma =$ CT, CC, TT, TC	CT, randomized to control, took treatment, etc.

randomized allocation so the effect of treatment in insistors can be absorbed into their base-line hazard $\lambda(t)\exp(\gamma_I)$. Also, no patients in class R receive treatment. We denote by $\{t_i\}$ the uncensored observation times arranged in increasing order, by $\{T_j\}$ all observation times (including censored observations) in increasing order and by $j \in R_i$ the fact that subject j is at risk at time t_i-, i.e. $T_j \geqslant t_i$. Note that there is no information about any effect of treatment in refusers or insistors. In particular, an interaction of treatment with insistor status is not identifiable and no assumptions about potential interactions are made.

3. Estimation

In this section we develop three estimators of increasing complexity, which are applicable to increasingly general situations.

3.1. No covariates

When there are no covariates a non-iterative generalization of the approach that is used in the binary case can be developed to yield a Mantel–Haenszel-type estimator. Let N_i^σ, $\sigma = \mathrm{CT}, \mathrm{CC}, \mathrm{TT}, \mathrm{TC}$, be the number of subjects at risk in each of the four observed groups at the time just before the ith failure and let n_{ki}^l be the (in some cases unobserved) number of individuals who are at risk just before the ith failure in the six groups that are indexed by $l = \mathrm{T}, \mathrm{C}$ (treated or control), and $k \equiv \mathrm{I}, \mathrm{A}, \mathrm{R}$ (insistors, ambivalent or refusers). Thus $N_i^{\mathrm{TT}} = n_{\mathrm{I}i}^{\mathrm{T}} + n_{\mathrm{A}i}^{\mathrm{T}}$, $N_i^{\mathrm{CT}} = n_{\mathrm{I}i}^{\mathrm{C}}$, etc. If D_i^σ is the indicator for the ith failure occurring in observed group σ, $\sigma = \mathrm{CT}$, CC, TT, TC, and $\Pr_i$ and E_i are respectively the conditional probability and expectation given the σ-field $\Im_{t_i-}$ of {all events up to (but not including) t_i and the fact that a failure occurred at time t_i}, then

$$\begin{aligned}\Pr_i(D_i^{\mathrm{TT}} = 1) = E_i(D_i^{\mathrm{TT}}) &= d_i\, \lambda(t_i)\, E\{n_{\mathrm{I}i}^{\mathrm{T}} \exp(\gamma_{\mathrm{I}}) + n_{\mathrm{A}i}^{\mathrm{T}} \exp(\gamma_{\mathrm{T}})\} \\ &= d_i\, \lambda(t_i)\, E\{n_{\mathrm{I}i}^{\mathrm{T}} \exp(\gamma_{\mathrm{I}}) + (N_i^{\mathrm{TT}} - n_{\mathrm{I}i}^{\mathrm{T}}) \exp(\gamma_{\mathrm{T}})\}\end{aligned}$$

where d_i is a normalizing constant such that $\Sigma_\sigma\, E_i(D_i^\sigma) = 1$. Now, if $\rho = (N^{\mathrm{TT}} + N^{\mathrm{TC}})/(N^{\mathrm{CT}} + N^{\mathrm{CC}})$ is the ratio of treated to control subjects at base-line, then $E_i(n_{\mathrm{I}i}^{\mathrm{T}}) = \rho N_i^{\mathrm{CT}}$, and

$$\Pr_i(D_i^{\mathrm{TT}} = 1) = E_i(D_i^{\mathrm{TT}}) = d_i\, \lambda(t_i)\{\rho N_i^{\mathrm{CT}} \exp(\gamma_{\mathrm{I}}) + (N_i^{\mathrm{TT}} - \rho N_i^{\mathrm{CT}}) \exp(\gamma_{\mathrm{T}})\}.$$

Also, since

$$\Pr_i(D_i^{\mathrm{CT}} = 1) = E_i(D_i^{\mathrm{CT}}) = d_i\, \lambda(t_i) N_i^{\mathrm{CT}} \exp(\gamma_{\mathrm{I}}),$$

we have

$$E_i\left(\frac{D_i^{\mathrm{TT}}-\rho D_i^{\mathrm{CT}}}{N_i^{\mathrm{TT}}-\rho N_i^{\mathrm{CT}}}\right)=d_i\,\lambda(t_i)\,\exp(\gamma_{\mathrm{T}}). \tag{2}$$

Similarly

$$\begin{aligned}E_i(D_i^{\mathrm{CC}})&=d_i\,\lambda(t_i)E_i\{n_{\mathrm{A}i}^{\mathrm{C}}+n_{\mathrm{R}i}^{\mathrm{C}}\,\exp(\gamma_{\mathrm{R}})\}\\&=d_i\,\lambda(t_i)\{N_i^{\mathrm{CC}}-\rho^{-1}N_i^{\mathrm{TC}}+\rho^{-1}N_i^{\mathrm{TC}}\,\exp(\gamma_{\mathrm{R}})\}\end{aligned}$$

so that

$$E_i\left(\frac{D_i^{\mathrm{CC}}-\rho^{-1}D_i^{\mathrm{TC}}}{N_i^{\mathrm{CC}}-\rho^{-1}N_i^{\mathrm{TC}}}\right)=d_i\,\lambda(t_i). \tag{3}$$

It follows that equation (2) divided by equation (3) equals $\exp(\gamma_{\mathrm{T}})$ and that for any (reasonable) set of $\Im_{t_i^-}$-measurable weights $\{w_i\}$

$$\hat{\theta}_{\mathrm{T}}=\frac{\sum_i w_i(D_i^{\mathrm{T}}/N_i^{\mathrm{T}})}{\sum_i w_i(D_i^{\mathrm{C}}/N_i^{\mathrm{C}})} \tag{4}$$

where

$$\begin{aligned}D_i^{\mathrm{T}}&=D_i^{\mathrm{TT}}-\rho D_i^{\mathrm{CT}},\\N_i^{\mathrm{T}}&=N_i^{\mathrm{TT}}-\rho N_i^{\mathrm{CT}},\\D_i^{\mathrm{C}}&=D_i^{\mathrm{CC}}-\rho^{-1}D_i^{\mathrm{TC}},\\N_i^{\mathrm{C}}&=N_i^{\mathrm{CC}}-\rho^{-1}N_i^{\mathrm{TC}}\end{aligned}$$

is an asymptotically unbiased estimate of the hazard ratio $\exp(\gamma_{\mathrm{T}})$ that is associated with the new treatment, and we omit terms in both the numerator and the denominator if either N_i^{T} or N_i^{C} is less than or equal to 0. This convention will be used in all expressions given below unless stated otherwise.

Similarly

$$\hat{\theta}_{\mathrm{I}}=\frac{\sum w_{\mathrm{I}i}(D_i^{\mathrm{CT}}/N_i^{\mathrm{CT}})}{\sum w_{\mathrm{I}i}(D_i^{\mathrm{C}}/N_i^{\mathrm{C}})},$$

and

$$\hat{\theta}_{\mathrm{R}}=\frac{\sum w_{\mathrm{R}i}(D_i^{\mathrm{TC}}/N_i^{\mathrm{TC}})}{\sum w_{\mathrm{R}i}(D_i^{\mathrm{C}}/N_i^{\mathrm{C}})}$$

are asymptotically unbiased estimates of $\theta_{\mathrm{I}}=\exp(\gamma_{\mathrm{I}})$ and $\theta_{\mathrm{R}}=\exp(\gamma_{\mathrm{R}})$ respectively, for weight systems $\{w_{\mathrm{I}i}\}$ and $\{w_{\mathrm{R}i}\}$ respectively. A step-by-step evaluation of these estimators is provided in Section 4.1.

3.1.1. A variance estimator

To calculate the asymptotic variance of $\log(\hat{\theta})$, note that, when the variances are small compared with the means for A and B and $\theta=E(A)/E(B)$,

$$\begin{aligned}\mathrm{var}\{\log(A/B)\}&\approx\mathrm{var}\{A/E(A)-B/E(B)\}\\&=\mathrm{var}(A-\theta B)/E^2(A).\end{aligned}$$

Conditional independence then implies that

$$\mathrm{var}\Big\{\sum_i w_i(D_i^{\mathrm{T}}/N_i^{\mathrm{T}} - \theta_{\mathrm{T}} D_i^{\mathrm{C}}/N_i^{\mathrm{C}})\Big\} = E\Big\{\sum_i w_i^2 \,\mathrm{var}(D_i^{\mathrm{T}}/N_i^{\mathrm{T}} - \theta_{\mathrm{T}} D_i^{\mathrm{C}}/N_i^{\mathrm{C}})\Big\}.$$

Expanding D_i^{T} and D_i^{C} and noting that $D_i^{\sigma} D_i^{\sigma'} = 0$ for $\sigma \neq \sigma' \in \{\mathrm{CT}, \mathrm{CC}, \mathrm{TT}, \mathrm{TC}\}$, we have

$$\mathrm{var}(D_i^{\mathrm{T}}/N_i^{\mathrm{T}} - \theta_{\mathrm{T}} D_i^{\mathrm{C}}/N_i^{\mathrm{C}}) = E\left\{\frac{D_i^{\mathrm{TT}} + \rho^2 D_i^{\mathrm{CT}}}{(N_i^{\mathrm{T}})^2} + \theta_{\mathrm{T}}^2 \frac{D_i^{\mathrm{CC}} + \rho^{-2} D_i^{\mathrm{TC}}}{(N_i^{\mathrm{C}})^2}\right\}.$$

Let

$$K_i = (n_{\mathrm{A}i}^{\mathrm{T}}\theta_{\mathrm{T}} + n_{\mathrm{I}i}^{\mathrm{T}}\theta_{\mathrm{I}} + n_{\mathrm{R}i}^{\mathrm{T}}\theta_{\mathrm{R}} + n_{\mathrm{I}i}^{\mathrm{C}}\theta_{\mathrm{I}} + n_{\mathrm{A}i}^{\mathrm{C}} + n_{\mathrm{R}i}^{\mathrm{C}}\theta_{\mathrm{R}})^{-1}.$$

Then $E_i(D_i^{\mathrm{TT}}) = K_i(n_{\mathrm{A}i}^{\mathrm{T}}\theta_{\mathrm{T}} + n_{\mathrm{I}i}^{\mathrm{T}}\theta_{\mathrm{I}})$, $E_i(D_i^{\mathrm{CT}}) = K_i n_{\mathrm{I}i}^{\mathrm{C}}\theta_{\mathrm{I}}$, $E_i(D_i^{\mathrm{CC}}) = K_i(n_{\mathrm{A}i}^{\mathrm{C}} + n_{\mathrm{R}i}^{\mathrm{C}}\theta_{\mathrm{R}})$ and $E_i(D_i^{\mathrm{TC}}) = K_i n_{\mathrm{R}i}^{\mathrm{T}}\theta_{\mathrm{R}}$. Hence the *conditional* variance, given $\Im_{t_i^-}$ and the n_{ki}, is equal to

$$K_i\left\{\frac{n_{\mathrm{A}i}^{\mathrm{T}}\theta_{\mathrm{T}} + n_{\mathrm{I}i}^{\mathrm{T}}\theta_{\mathrm{I}} + \rho^2 n_{\mathrm{I}i}^{\mathrm{C}}\theta_{\mathrm{I}}}{(N_i^{\mathrm{T}})^2} + \theta_{\mathrm{T}}^2 \frac{n_{\mathrm{A}i}^{\mathrm{C}} + n_{\mathrm{R}i}^{\mathrm{C}}\theta_{\mathrm{R}} + \rho^{-2} n_{\mathrm{R}i}^{\mathrm{T}}\theta_{\mathrm{R}}}{(N_i^{\mathrm{C}})^2}\right\}.$$

Substituting the estimators $\hat{\theta}_j$ for the θ_j and replacing $n_{\mathrm{A}i}^{\mathrm{T}}$ by N_i^{T} and $n_{\mathrm{I}i}^{\mathrm{T}}$ by ρN_i^{CT}, etc. we have

$$\begin{aligned} K_i &\left\{\frac{N_i^{\mathrm{T}}\hat{\theta}_{\mathrm{T}} + \rho N_i^{\mathrm{CT}}\hat{\theta}_{\mathrm{I}} + \rho^2 N_i^{\mathrm{CT}}\hat{\theta}_{\mathrm{I}}}{(N_i^{\mathrm{T}})^2} + \hat{\theta}_{\mathrm{T}}^2 \frac{N_i^{\mathrm{C}} + \rho^{-1} N_i^{\mathrm{TC}}\hat{\theta}_{\mathrm{R}} + \rho^{-2} N_i^{\mathrm{TC}}\hat{\theta}_{\mathrm{R}}}{(N_i^{\mathrm{C}})^2}\right\} \\ &= \frac{\hat{\theta}_{\mathrm{T}} K_i}{N_i^{\mathrm{T}} N_i^{\mathrm{C}}}\left[N_i^{\mathrm{C}}\left\{1 + \rho(1+\rho)\frac{N_i^{\mathrm{CT}}\hat{\theta}_{\mathrm{I}}}{N_i^{\mathrm{T}}\hat{\theta}_{\mathrm{T}}}\right\} + \hat{\theta}_{\mathrm{T}} N_i^{\mathrm{T}}\left\{1 + \rho^{-1}(1+\rho^{-1})\frac{N_i^{\mathrm{TC}}\hat{\theta}_{\mathrm{R}}}{N_i^{\mathrm{C}}}\right\}\right] \\ &= \hat{\theta}_{\mathrm{T}} K_i W_i \qquad \text{(say)}. \end{aligned}$$

Similarly we can approximate $E_i(D_i^{\mathrm{T}}/N_i^{\mathrm{T}})$ by $K_i\hat{\theta}_{\mathrm{T}}$. Thus the variance of $\log(\hat{\theta}_{\mathrm{T}})$ using weights $\{w_i\}$ is approximately

$$\widehat{\mathrm{var}}\{\log(\hat{\theta}_{\mathrm{T}})\} = \Big(\sum_i w_i^2 \hat{K}_i W_i\Big) \Big/ \Big(\sum_i w_i \hat{K}_i\Big)^2 \hat{\theta}_{\mathrm{T}}, \tag{5}$$

where

$$\hat{K}_i = \{N_i^{\mathrm{T}}\hat{\theta}_{\mathrm{T}} + (1+\rho)N_i^{\mathrm{CT}}\hat{\theta}_{\mathrm{I}} + N_i^{\mathrm{C}} + (1+\rho^{-1})N_i^{\mathrm{TC}}\hat{\theta}_{\mathrm{R}}\}^{-1}.$$

3.1.2. Choice of weights

It remains to choose an appropriate weight system for equation (4). One possible choice which is not too inefficient when $\theta = 1$ and $\rho = 1$ is $w_i = N_i^{\mathrm{T}} N_i^{\mathrm{C}}/(N_i^{\mathrm{T}} + N_i^{\mathrm{C}})$ which gives the estimates

$$\tilde{\theta}_{\mathrm{MH}} = \frac{\sum_i D_i^{\mathrm{T}} N_i^{\mathrm{C}}/(N_i^{\mathrm{C}} + N_i^{\mathrm{T}})}{\sum_i D_i^{\mathrm{C}} N_i^{\mathrm{T}}/(N_i^{\mathrm{C}} + N_i^{\mathrm{T}})}$$

(where the sums are only for values of i for which $N_i^{\mathrm{C}} + N_i^{\mathrm{T}} > 0$). This looks like the classical Mantel–Haenszel estimator and in fact it reduces to the classical Mantel–Haenszel estimate in the absence of contamination or non-compliance. More generally, the observed numbers that are used by the Mantel–Haenszel estimator are replaced by the estimates of these (now unobserved)

quantities for the ambivalent group. When θ differs substantially from 1, a one-step estimator can be obtained but this requires preliminary estimates of θ_I and θ_R (McKeague and Sasieni, 1994).

Efficient weights are given by $w_i = 1/W_i$ and in this case the asymptotic variance estimator simplifies to

$$\widehat{\text{var}}\{\log(\hat{\theta}_T)\} = \left(\hat{\theta}_T \sum \hat{K}_i / W_i\right)^{-1}.$$

3.2. Covariates which are independent of class membership

When covariates exist, it is not possible to use the Mantel–Haenszel approach (except where the covariates take just a few values, so that a stratified approach can be used). In this case a full likelihood can be written down, but it is tedious to solve, and the partial likelihood that is developed below is more tractable.

This method is based on the martingale approach to counting processes, which has its roots in the partial likelihood that was developed by Cox (1975) for the classical proportional hazards model. The major novelty is that membership in the three unobserved classes (I, A and R) is re-evaluated at each failure time t_i, based only on the observed set {CT, CC, TT, TC} to which individuals in the risk set R_i belong at this time. A requirement of this model is that the covariates are independent of the underlying class membership (insistors, ambivalent and refusers).

This leads to consideration of the likelihood

$$L = \prod_i \left(\frac{\theta_i}{\sum_{j \in R_i} \theta_{ji}} \right)^{\delta_i}, \tag{6}$$

where the hazard ratios θ_{ji} are given by

$$\theta_{ji} = \begin{cases} \exp(\beta z_j + \gamma_I), & j \in \text{CT}, \\ \exp(\beta z_j)\{1 - \pi_{Ri} + \pi_{Ri} \exp(\gamma_R)\}, & j \in \text{CC}, \\ \exp(\beta z_j)\{\pi_{Ii} \exp(\gamma_I) + (1 - \pi_{Ii}) \exp(\gamma_T)\}, & j \in \text{TT}, \\ \exp(\beta z_j + \gamma_R), & j \in \text{TC}, \end{cases}$$

and $\theta_i = \theta_{ii}$. The π_{ki} are based on the composition of the four groups at time t_i, i.e.

$$\begin{aligned} \pi_{Ii} &= \min\left(\frac{N_i^{\text{CT}}}{N_i^{\text{TT}}} \rho, 1 \right), \\ \pi_{Ri} &= \min\left(\frac{N_i^{\text{TC}}}{N_i^{\text{CC}}} \rho^{-1}, 1 \right). \end{aligned} \tag{7}$$

The π_{Ii} can be interpreted as estimates of the proportion of individuals complying with randomization to treatment and at risk at times t_i- who would have insisted on having the new treatment even if they had been randomized to control. The π_{Ri} have a similar interpretation as the estimated proportion of refusers among the CC group at t_i.

This approach allows a much easier analysis than the full likelihood since the π_{ki} do not depend on the unknown parameters (β, γ) or the base-line hazard function. Also it directly makes use of the conditional status of the risk sets at time t_i. It is worth noting that randomization only guarantees approximately the same proportion of indifferent patients in each treatment arm at base-line. If there is a treatment effect, this will not be so subsequently. However, randomization guarantees approximately the same proportion of insistors and refusers in each arm at all follow-up times, and that they will have approximately the same covariate distribu-

tion at all times, whereas before we have assumed that class membership is independent of all covariates.

Familiar calculations yield the score functions

$$\frac{\mathrm{d}\{\log(L)\}}{\mathrm{d}\gamma_k} = \sum_i \delta_i (v_{ki} - \overline{v_{ki}}), \qquad k = \mathrm{T}, \mathrm{I}, \mathrm{R},$$
$$\frac{\mathrm{d}\{\log(L)\}}{\mathrm{d}\beta} = \sum_i \delta_i (z_i - \bar{z}_i), \tag{8}$$

where

$$v_{kji} = \begin{cases} \dfrac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\gamma_\mathrm{T}} = \dfrac{(1-\pi_{\mathrm{I}i})\exp(\gamma_\mathrm{T})}{\pi_{\mathrm{I}i}\exp(\gamma_\mathrm{I}) + (1-\pi_{\mathrm{I}i})\exp(\gamma_\mathrm{T})} I_{j\in\mathrm{TT}}, & k = \mathrm{T}, \\ \dfrac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\gamma_\mathrm{I}} = I_{j\in\mathrm{CT}} + \dfrac{\pi_{\mathrm{I}i}\exp(\gamma_\mathrm{I})}{\pi_{\mathrm{I}i}\exp(\gamma_\mathrm{I}) + (1-\pi_{\mathrm{I}i})\exp(\gamma_\mathrm{T})} I_{j\in\mathrm{TT}}, & k = \mathrm{I}, \\ \dfrac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\gamma_\mathrm{R}} = I_{j\in\mathrm{TC}} + \dfrac{\pi_{\mathrm{R}i}\exp(\gamma_\mathrm{R})}{\pi_{\mathrm{R}i}\exp(\gamma_\mathrm{R}) + (1-\pi_{\mathrm{R}i})} I_{j\in\mathrm{CC}}, & k = \mathrm{R}, \end{cases}$$

$v_{ki} \equiv v_{kii}$, and for a general $x \equiv (x_1, \ldots, x_I)$, x_j a scalar, vector or matrix

$$\bar{x}_i = \sum_{j\in R_i} x_j \theta_{ji} \Big/ \sum_{j\in R_i} \theta_{ji}.$$

Note that the v_{kji} are estimates of the probability that an individual has hazard ratio $\exp(\gamma_k)$ ($k \in \{\mathrm{T}, \mathrm{I}, \mathrm{R}\}$) given their randomized and actual treatment status and that they are at risk at time t_i-. A simple example is evaluated in Section 4.1.

3.2.1. Variance estimator

Because of the form of the likelihood (6), the Hessian takes the special form

$$\frac{\mathrm{d}^2\{\log(L)\}}{\mathrm{d}m_k\,\mathrm{d}m_l} = \sum_i \delta_i \left(\frac{\mathrm{d}^2\{\log(\theta_{ii})\}}{\mathrm{d}m_k\,\mathrm{d}m_l} - \frac{\sum_{j\in R_i}\theta_{ji}\,\mathrm{d}^2\{\log(\theta_{ji})\}/\mathrm{d}m_k\,\mathrm{d}m_l}{\sum_{j\in R_i}\theta_{ji}} - \left[\frac{\sum_{j\in R_i}\theta_{ji}[\mathrm{d}\{\log(\theta_{ji})\}/\mathrm{d}m_k]\,\mathrm{d}\{\log(\theta_{ji})\}/\mathrm{d}m_l}{\sum_{j\in R_i}\theta_{ji}} - \frac{\sum_{j\in R_i}\theta_{ji}\,\mathrm{d}\{\log(\theta_{ji})\}/\mathrm{d}m_k}{\sum_{j\in R_i}\theta_{ji}} \frac{\sum_{j\in R_i}\theta_{ji}\,\mathrm{d}\{\log(\theta_{ji})\}/\mathrm{d}m_l}{\sum_{j\in R_i}\theta_{ji}} \right] \right)$$

for general variables (m_k, m_l). When the parameters are equal to their true values, the second term is the conditional expectation of the first term given the risk set at time t_i. It follows that, when divided by the number of events, the first two terms will tend to 0. The remaining two terms are in the form of (minus) a sum of covariance matrices, and thus will always be negative definite. Thus, regardless of the form of θ_{ji}, at the true parameter value, the likelihood will asymptotically be concave under very general conditions.

In our case the terms of the information matrix are given by the following specific formulae:

$$\left.\begin{aligned} \frac{\mathrm{d}^2\{\log(L)\}}{\mathrm{d}\gamma_k\,\mathrm{d}\gamma_l} &= \sum_i \delta_i\{\Delta_{kl}(v_{ki}-\overline{v_{ki}}) - (v_{ki}v_{li}-\overline{v_{ki}}\,\overline{v_{li}})\}, && k,l=\mathrm{T,\,I,\,R},\\ \frac{\mathrm{d}^2\{\log(L)\}}{\mathrm{d}\beta\,\mathrm{d}\beta^{\mathrm{T}}} &= -\sum_i \delta_i\{(\overline{zz^{\mathrm{T}}})_i - \bar{z}_i\bar{z}_i^{\mathrm{T}}\}, &&\\ \frac{\mathrm{d}^2\{\log(L)\}}{\mathrm{d}\beta\,\mathrm{d}\gamma_k} &= -\sum_i \delta_i\{(\overline{zv_k})_i - \bar{z}_i\overline{v_{ki}}\}, && k=\mathrm{T,\,I,\,R}, \end{aligned}\right\} \tag{9}$$

where Δ_{kl} is Kronecker's delta. Note that $v_{ki}v_{li}=0$ if $(k,l)=(\mathrm{T,\,R})$ or $(k,l)=(\mathrm{I,\,R})$. Since equation (6) is not a true likelihood, inversion of expression (9) will not always give the correct asymptotic variance.

3.2.2. *Asymptotic properties*

The asymptotic properties of the estimator that is defined by equating the score functions (8) to 0 can be summarized by the following theorem.

Theorem 1. Assume that the covariates z_i are bounded, and that the observations $(t_i, \delta_i, k_i, z_{0i}, z_i)$, $i=1,\ldots,n$, are independent and obey the model given by expression (1). Then, for sufficiently large n, there is a zero $(\hat{\gamma},\hat{\beta})$ of expression (8) which is a local maximum and is $n^{1/2}$ consistent for the true values (γ,β) and asymptotically normal. When $\gamma_{\mathrm{I}}=\gamma_{\mathrm{T}}$ and $\gamma_{\mathrm{R}}=0$, the information matrix can be approximated by expression (9) in the sense that the inverse of the covariance matrix of $n^{-1/2}(\gamma-\hat{\gamma},\beta-\hat{\beta})$ is asymptotically equivalent to n^{-1} times expression (9). More generally the covariance matrix can be obtained by empirical process methods from a first-order Taylor series expansion of the empirical survival processes for the four observed groups TT, TC, CT and CC as detailed below.

Proof. The existence of a solution $(\hat{\gamma},\hat{\beta})$ near the true parameter values is guaranteed by the asymptotic concavity of equation (6) at (γ,β). Some considerable calculations show that expression (8) is asymptotically linear at (γ,β) so expression (8) can be asymptotically approximated by a one-term Taylor series expansion. A difficulty arises since the $\pi_{\mathrm{I}i}$ are random variables and cannot be treated as constants. Furthermore, they are not conditional probabilities with respect to $\Im_{t_i-}$ so expression (8) is not quite a martingale at the true (γ,β,λ). This problem can be overcome by expanding the π_{ki} in those θ_{kij} that are implicit in the $\overline{v_{ki}}$ (but not in the v_{kij} themselves) in expression (8) about the true conditional probabilities and then using methods for empirical processes (see for example chapter 3 of Shorack and Wellner (1986)). Define $\bar{\pi}_{\mathrm{I}i}$ to be the conditional probability of being an insistor given $\Im_{t_i-}$ and that the individual is in group TT and is at risk at t_i- and define $\bar{\pi}_{\mathrm{R}i}$ analogously. Then

$$\begin{aligned} \bar{\pi}_{\mathrm{I}i} &= (N_i^{\mathrm{TT}})^{-1} \sum_{j\in R_i\cap \mathrm{TT}} \pi_{\mathrm{I}} S_{ji}^{\mathrm{I}}/\{\pi_{\mathrm{I}} S_{ji}^{\mathrm{I}} + (1-\pi_{\mathrm{I}})S_{ji}^{\mathrm{A}}\},\\ \bar{\pi}_{\mathrm{R}i} &= (N_i^{\mathrm{CC}})^{-1} \sum_{j\in R_i\cap \mathrm{CC}} \pi_{\mathrm{R}} S_{ji}^{\mathrm{R}}/\{\pi_{\mathrm{R}} S_{ji}^{\mathrm{R}} + (1-\pi_{\mathrm{R}})S_{ji}^{\mathrm{A}}\}, \end{aligned} \tag{10}$$

where S_{ji}^k is the survival function for individual with covariates z_j at t_i given the latent class, given by

$$S_{ji}^{k} \equiv \exp\{-\tau_{kj}\,\Lambda(t_i)\},$$
$$\tau_{kj} \equiv \exp(\gamma_k + \beta z_j + \gamma_{\mathrm{T}} z_{0j}), \qquad k = \mathrm{I, A, R}, \tag{11}$$

and the π_k are the π_{ki} that are determined at base-line, i.e. by using the N^{σ} at time 0. The main term for the covariance matrix of the score functions is given by the inverse of expression (9), as can be seen by martingale arguments. Some algebra shows that the first-order term in the Taylor series expansion is asymptotically independent of the main term and gives an additive component to the variance. In particular, the extra component to the covariance matrix of the score function (8) at the true parameter values takes the form

$$\mathrm{var}\Big[\sum_i \delta_i \{A_{ki}^{\mathrm{I}}(\pi_{\mathrm{I}i} - \bar{\pi}_{\mathrm{I}i}) + A_{ki}^{\mathrm{R}}(\pi_{\mathrm{R}i} - \bar{\pi}_{\mathrm{R}i})\}\Big],$$

where

$$A_{ki}^{l} = \left(\overline{v_{kji}\frac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\pi_{li}}}\right)_i - \overline{v_{ki}}\left(\overline{\frac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\pi_{li}}}\right)_i, \qquad l = \mathrm{I, R}, \quad k = \mathrm{T, I, R},$$

$$\frac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\pi_{\mathrm{I}i}} = \frac{\exp(\gamma_{\mathrm{I}}) - \exp(\gamma_{\mathrm{T}})}{\pi_{\mathrm{I}i}\exp(\gamma_{\mathrm{I}}) + (1 - \pi_{\mathrm{I}i})\exp(\gamma_{\mathrm{T}})} I_{j\in \mathrm{TT}}$$

and

$$\frac{\mathrm{d}\{\log(\theta_{ji})\}}{\mathrm{d}\pi_{\mathrm{R}i}} = \frac{\exp(\gamma_{\mathrm{R}}) - 1}{1 - \pi_{\mathrm{R}i} + \pi_{\mathrm{R}i}\exp(\gamma_{\mathrm{R}})} I_{j\in \mathrm{CC}}.$$

Similar expressions (denoted A_{zi}^{l}) arise for the covariates except that z_{ji} replaces v_{kji}.

When $\gamma_{\mathrm{I}} = \gamma_{\mathrm{T}}$ and $\gamma_{\mathrm{R}} = 0$ the first-order term in the expansion of the π_{ki} about the $\bar{\pi}_{ki}$ is 0 and can be ignored. In this case the martingale central limit theorem can be used to establish the asymptotic normality of the main term and to show that its asymptotic variance is approximated by expression (9), the second derivative with respect to the (γ, β) of the score function (with $\pi_{\mathrm{I}i}$ replaced by $\bar{\pi}_{\mathrm{I}i}$). Furthermore, n^{-1} times this expression can be seen to approximate a constant given by the expected information, so n times the covariance matrix of $(\hat{\gamma} - \gamma, \hat{\beta} - \beta)$ will be given by the inverse of n^{-1} times expression (9). More generally (i.e. when $\gamma_{\mathrm{I}} \neq \gamma_{\mathrm{T}}$ or $\gamma_{\mathrm{R}} \neq 0$), the derivative term with respect to the π_{ki} cannot be ignored and the use of expression (9) to estimate the variance will lead to an underestimate.

Now, from expression (7), $\pi_{\mathrm{I}i}$ and $\pi_{\mathrm{R}i}$ are both ratios of empirical decrement processes (for groups CT, TT, CC and TC) truncated at 1, and from expression (10) $\bar{\pi}_{\mathrm{I}i}$ and $\bar{\pi}_{\mathrm{R}i}$ are ratios of weighted empirical processes based on elements of groups TT and CC respectively. All four processes (N^{σ}; $\sigma = \mathrm{CT, TC, CC, TT}$) are independent, so asymptotically

$$\begin{aligned}
\pi_{\mathrm{I}i} - \bar{\pi}_{\mathrm{I}i} &\cong \rho\Big\{\sum_{l\in \mathrm{CT}}(I_{li} - S_{li})\Big\}/S_i^{\mathrm{TT}} - \sum_{l\in \mathrm{TT}}\{\rho S_i^{\mathrm{CT}}/(S_i^{\mathrm{TT}})^2 + (a_{li}^{\mathrm{I}} - \bar{\pi}_{\mathrm{I}i})/N_i^{\mathrm{TT}}\}(I_{li} - S_{li}), \\
\pi_{\mathrm{R}i} - \bar{\pi}_{\mathrm{R}i} &\cong \rho^{-1}\Big\{\sum_{l\in \mathrm{TC}}(I_{li} - S_{li})\Big\}/S_i^{\mathrm{CC}} - \sum_{l\in \mathrm{CC}}\{\rho^{-1} S_i^{\mathrm{TC}}/(S_i^{\mathrm{CC}})^2 + (a_{li}^{\mathrm{R}} - \bar{\pi}_{\mathrm{R}i})/N_i^{\mathrm{CC}}\}(I_{li} - S_{li}),
\end{aligned} \tag{12}$$

where $I_{li} \equiv I_{\{T_l \geqslant t_i\}}$, $S_i^{\sigma} \equiv \Sigma_{l\in\sigma}\, S_{li}$, $\sigma \in \{\mathrm{CT, CC, TT, TC}\}$,

$$S_{li} = E(I_{li}) = \Pr(T_l \geqslant t_i) = \begin{cases} S_{li}^{\mathrm{I}}, & l \in \mathrm{CT}, \\ \pi_{\mathrm{I}} S_{li}^{\mathrm{I}} + (1-\pi_{\mathrm{I}}) S_{li}^{\mathrm{A}}, & l \in \mathrm{TT}, \\ \pi_{\mathrm{R}} S_{li}^{\mathrm{R}} + (1-\pi_{\mathrm{R}}) S_{li}^{\mathrm{A}}, & l \in \mathrm{CC}, \\ S_{li}^{\mathrm{R}}, & l \in \mathrm{TC}, \end{cases}$$

and the S_{li}^k are given in expression (11). Also,

$$a_{li}^k = \pi_k S_{li}^k / \{\pi_k S_{li}^k + (1-\pi_k) S_{li}^A\}, \qquad k = \mathrm{I}, \mathrm{R},$$

denotes the individual summands that are used in expression (10). Note that the left-hand sides in the two equations (12) are independent.

If we replace the A_{ki}^l by their expectations, evaluate the variances and then replace the expectations by their observed values (effectively treat them as constants) we arrive at the following approximation for the covariance matrix between the scores for the (γ_k, z):

$$\sum_i \sum_j \delta_i \delta_j \{A_{ki}^{\mathrm{I}} A_{lj}^{\mathrm{I}} \operatorname{cov}(\pi_{\mathrm{I}i}, \pi_{\mathrm{I}j}) + A_{ki}^{\mathrm{R}} A_{lj}^{\mathrm{R}} \operatorname{cov}(\pi_{\mathrm{R}i}, \pi_{\mathrm{R}j})\}, \qquad l \equiv \mathrm{T}, \mathrm{I}, \mathrm{R}, z.$$

Standard calculations for marked empirical processes of independent but not identically distributed random variables using expression (12) give that asymptotically for $t_i \geqslant t_j$

$$\begin{aligned} \operatorname{cov}(\pi_{\mathrm{I}i}, \pi_{\mathrm{I}j}) = \rho^2 \Big\{ \sum_{l \in \mathrm{CT}} S_{li}(1-S_{lj}) \Big\} \Big/ S_i^{\mathrm{TT}} S_j^{\mathrm{TT}} + \sum_{l \in \mathrm{TT}} \{\rho S_i^{\mathrm{CT}} / (S_i^{\mathrm{TT}})^2 + (a_{li}^{\mathrm{I}} - \bar{\pi}_{\mathrm{I}i}) / N_i^{\mathrm{TT}}\} \\ \times \{\rho S_j^{\mathrm{CT}} / (S_j^{\mathrm{TT}})^2 + (a_{lj}^{\mathrm{I}} - \bar{\pi}_{\mathrm{I}j}) / N_j^{\mathrm{TT}}\} S_{li}(1-S_{lj}), \\ \operatorname{cov}(\pi_{\mathrm{R}i}, \pi_{\mathrm{R}j}) = \rho^{-2} \Big\{ \sum_{l \in \mathrm{TC}} S_{li}(1-S_{lj}) \Big\} \Big/ S_i^{\mathrm{CC}} S_j^{\mathrm{CC}} + \sum_{l \in \mathrm{CC}} \{\rho S_i^{\mathrm{TC}} / (S_i^{\mathrm{CC}})^2 + (a_{li}^{\mathrm{R}} - \bar{\pi}_{\mathrm{R}i}) / N_i^{\mathrm{CC}}\} \\ \times \{\rho^{-1} S_j^{\mathrm{TC}} / (S_j^{\mathrm{CC}})^2 + (a_{lj}^{\mathrm{R}} - \bar{\pi}_{\mathrm{R}j}) / N_j^{\mathrm{CC}}\} S_{li}(1-S_{lj}). \end{aligned}$$

Computation of the variance requires an estimate of the base-line cumulative hazard in expression (11). In keeping with the partial likelihood concept, we can use the π_{ki} and estimates of the (β, γ) to obtain the estimate

$$\hat{\Lambda}(t_i) = \sum_{j \leqslant i} \delta^j \Delta_j,$$

where, when t_j is uncensored, $\Delta_j^{-1} = \Sigma_{l \in R_j} \theta_{lj}$. Thus if we let C denote the usual covariance for the scores that are given by expression (9) and D the extra term that is associated with the variability of the π_{ki}, the total covariance is $C + D$ and the usual Taylor series expansion gives

$$C^{-1} - C^{-1} D C^{-1} \tag{13}$$

as the asymptotic covariance for the estimates (γ, β). This is valid when all the eigenvalues of $C^{-1}D$ are small.

3.3. Covariates correlated with class membership

In the previous section we needed to assume that covariates were independent of class membership for the proportions $\pi_{\mathrm{I}i}$ and $\pi_{\mathrm{R}i}$ to be valid at follow-up times. When class membership does depend on covariates, $\pi_{\mathrm{I}i}$ and $\pi_{\mathrm{R}i}$ will also depend on the covariate values for t_i. In particular, the $\{\pi_{ki}\}$ need to be modified to reflect this by replacing $\pi_{\mathrm{I}i}$ by $\Pr(i \in \text{insistors} | z_i, i \in \mathrm{TT} \cap \mathrm{R}i)$ and $\pi_{\mathrm{R}i}$ by $\Pr(i \in \text{refusers} | z_i, i \in \mathrm{CC} \cap \mathrm{R}i)$. Separate regressions can be used to test whether such

effects exist and to adjust for them if necessary. Again the control group is used to examine insistors and the treated group is used for refusers.

In the general case some sort of smoothing of the covariate distribution for the three classes at time t_i is needed, and the π_{ki} are computed locally in a neighbourhood of z_i. This requires a large data set and may not be practical in many situations.

The scope for parametric simplification of the covariate distribution is limited by the general form of the survival model. In particular if $F_k(z)$ is the distribution of covariates in class k at base-line, the theoretical distribution at time t_i will be of the form

$$\lambda^{\delta_i}\tau_{ki}^{\delta_i} S_i^k \, \mathrm{d}F_k(z)$$

where τ_{ki} and $S_i^k = S_{ii}^k$ are defined in expression (10), so any parametric model of the relationship between the covariate distributions in different classes at time $\{t_i\}$ must involve the base-line cumulative hazard function $\Lambda(t)$ unless no survival effects are associated with class membership (i.e. $\gamma_{\mathrm{I}} = \gamma_{\mathrm{T}}$ and $\gamma_{\mathrm{R}} = 0$). Avoidance of having to compute this was the main reason for using the partial likelihood in the first place, so, when the covariates are related to class membership, it is probably best to resort to the full likelihood, where we need only to specify the relationship between covariates and class membership at base-line.

3.3.1. Full likelihood

Everyone in group CT is an insistor and everyone in group TC is a refuser. Hence the likelihood in these groups takes the usual form. Individuals in group CC are either ambivalent or refusers and so the likelihood is a mixture of the two likelihoods. The full likelihood takes the form of a product of the following terms:

$$L_i = \begin{cases} \lambda^{\delta_i}(t_i)\, p_i^{\mathrm{I}}, & i \in \mathrm{CT}, \\ \lambda^{\delta_i}(t_i)[\{1-\pi_{\mathrm{R}}(z)\}\, p_i^{\mathrm{A}} + \pi_{\mathrm{R}}(z)\, p_i^{\mathrm{R}}], & i \in \mathrm{CC}, \\ \lambda^{\delta_i}(t_i)[\pi_{\mathrm{I}}(z)\, p_i^{\mathrm{I}} + \{1-\pi_{\mathrm{I}}(z)\}\, p_i^{\mathrm{A}}], & i \in \mathrm{TT}, \\ \lambda^{\delta_i}(t_i)\, p_i^{\mathrm{R}}, & i \in \mathrm{TC}, \end{cases}$$

where, for $k = \mathrm{I}, \mathrm{A}, \mathrm{R}$, $p_i^k \equiv \tau_{ki}^{\delta_i} S_i^k$ and τ_{ki} and $S_i^k \equiv S_{ii}^k$ are defined in expression (11), and $\pi_k(z)$ determines the probability distribution of class membership at base-line of an individual with covariates z in observed group CC ($k = \mathrm{R}$) or TT ($k = \mathrm{I}$). When class membership does not depend on covariates the π_k are just the π_{ki}, determined at base-line. More generally, this could be done by estimating the conditional probability from smoothed densities. An alternative approach is to assume that the base-line covariate densities belong to an exponential family, i.e. $\mathrm{d}F_k(z) = \exp(\mu_k z)\, \mathrm{d}F_{\mathrm{A}}(z)$, $k = \mathrm{I}, \mathrm{R}$. In this case the π_k are replaced by

$$\frac{\pi_k \exp(\mu_k z)}{\pi_k \exp(\mu_k z) + 1 - \pi_k}, \qquad k = \mathrm{I}, \mathrm{R}.$$

The values of μ_{I} and μ_{R} can be estimated as follows.

At base-line we have, in an obvious notation,

$$F_{\mathrm{TT}}(z) = \pi_{\mathrm{I}}\, F_{\mathrm{I}}(z) + (1-\pi_{\mathrm{I}})\, F_{\mathrm{A}}(z)$$

and, since $F_{\mathrm{CT}} = F_{\mathrm{I}}$ by randomization, we seek μ_{I} so that

$$d(F_{\mathrm{TT}} - \pi_{\mathrm{I}} F_{\mathrm{CT}})/(1-\pi_{\mathrm{I}}) = \exp(\mu_{\mathrm{I}} z)\, \mathrm{d}F_{\mathrm{CT}}(z) \Big/ \int \exp(\mu_{\mathrm{I}} z)\, \mathrm{d}F_{\mathrm{CT}}(z).$$

The value of μ_I can then be estimated by determining when the associated empirical distributions are approximately equal. The simplest approach is to equate means, i.e. to solve

$$(N_0^T)^{-1}\Big(\sum_{i\in TT} z_i - \rho \sum_{i\in CT} z_i\Big) = \sum_{i\in CT} z_i \exp(\mu_I z_i)\Big/ \sum_{i\in CT} \exp(\mu_I z_i) \tag{14}$$

for μ_I. More robust methods using trimmed means, medians, etc. can also be used if necessary when outliers are present. Similarly we can solve for μ_R from

$$(N_0^C)^{-1}\Big(\sum_{i\in CC} z_i - \rho^{-1} \sum_{i\in TC} z_i\Big) = \sum_{i\in TC} z_i \exp(\mu_R z_i)\Big/ \sum_{i\in TC} \exp(\mu_R z_i). \tag{15}$$

Further, since the covariate distributions of treated and control ambivalent patients are the same owing to randomization, we can combine these groups and replace the left-hand sides of equations (14) and (15) by

$$(N_0^C + N_0^T)^{-1}\Big(\sum_{i\in CC} z_i + \sum_{i\in TT} z_i - \sum_{i\in CT} \rho z_i - \sum_{i\in TC} \rho^{-1} z_i\Big).$$

The existence of a unique solution for μ_I and μ_R even in the vector case is guaranteed provided that $N_0^T > 0$, $N_0^C > 0$ and the covariate distribution is not degenerate, since then the derivative of the right-hand side of equations (14) and (15) with respect to μ_I and μ_R respectively is positive definite.

3.3.2. *Estimation of the main parameters*

To ease notation we order the t_i in increasing order. As usual we obtain estimating equations from the derivatives of the log-likelihood:

$$\begin{aligned}
\frac{d\{\log(L)\}}{d\gamma_I} &= \sum_{i\in CT} \{\delta_i - \exp(\gamma_I + \beta z_i)\,\Lambda(t_i)\} + \sum_{i\in TT} \{\delta_i - \exp(\gamma_I + \beta z_i)\,\Lambda(t_i)\} \\
&\quad \times \frac{\pi_I(z_i) p_i^I}{\pi_I(z_i) p_i^I + \{1-\pi_I(z_i)\} p_i^A} \\
\frac{d\{\log(L)\}}{d\gamma_R} &= \sum_{i\in TC} \{\delta_i - \exp(\gamma_R + \beta z_i)\,\Lambda(t_i)\} + \sum_{i\in CC} \{\delta_i - \exp(\gamma_R + \beta z_i)\,\Lambda(t_i)\} \\
&\quad \times \frac{\pi_R(z_i) p_i^R}{\{1-\pi_R(z_i)\} p_i^A + \pi_R(z_i) p_i^R} \\
\frac{d\{\log(L)\}}{d\gamma_T} &= \sum_{i\in TT} \{\delta_i - \exp(\gamma_T + \beta z_i)\,\Lambda(t_i)\} \frac{\{1-\pi_I(z_i)\} p_i^A}{\pi_I(z_i) p_i^I + \{1-\pi_I(z_i)\} p_i^A},
\end{aligned}$$

and for the covariates $l = 1, \ldots, m$

$$\begin{aligned}
\frac{d\{\log(L)\}}{d\beta_l} = \sum_i z_{li} \Bigg\{ \delta_i - \exp(\beta z_i)\,\Lambda(t_i) \Bigg(& \exp(\gamma_I) \left[I_{i\in CT} + \frac{\pi_I(z_i) p_i^I}{\pi_I(z_i) p_i^I + \{1-\pi_I(z_i)\} p_i^A} I_{i\in TT} \right] \\
& + \frac{\{1-\pi_R(z_i)\} p_i^A}{\{1-\pi_R(z_i)\} p_i^A + \pi_R(z_i) p_i^R} I_{i\in CC} + \exp(\gamma_T) \frac{\{1-\pi_I(z_i)\} p_i^A}{\pi_I(z_i) p_i^I + \{1-\pi_I(z_i)\} p_i^A} I_{i\in TT} \\
& + \exp(\gamma_R) \left[I_{i\in TC} + \frac{\pi_R(z_i) p_i^R}{\{1-\pi_R(z_i)\} p_i^A + \pi_R(z_i) p_i^R} I_{i\in CC} \right] \Bigg) \Bigg\}.
\end{aligned}$$

These are not yet useful estimating equations, because $\Lambda(t_i)$ is unknown. We follow the conventional approach (Breslow, 1974; Kalbfleisch and Prentice, 1980) of replacing $\Lambda(t_i)$ by the discrete maximum likelihood estimate, i.e. assume that $\Lambda(t_i) = \Sigma_{t_j \leqslant t_i} \Delta_j$ and replace $\lambda^{\delta_i}(t_i)$ by $\Delta_i^{\delta_i}$ in the likelihood. We then form a profile likelihood by estimating Δ_i for fixed values of (β, γ). Standard calculations yield

$$\frac{\mathrm{d}\{\log(L)\}}{\mathrm{d}\Delta_i} = \frac{\delta_i}{\Delta_i} - \left[\exp(\gamma_{\mathrm{I}}) \sum_{\substack{j \in \mathrm{CT} \\ j \geqslant i}} \exp(\beta z_j) + \exp(\gamma_{\mathrm{R}}) \sum_{\substack{j \in \mathrm{TC} \\ j \geqslant i}} \exp(\beta z_j) + \sum_{\substack{j \in \mathrm{CC} \\ j \geqslant i}} \exp(\beta z_j) \frac{\{1 - \pi_{\mathrm{R}}(z_j)\} p_j^{\mathrm{A}} + \exp(\gamma_{\mathrm{R}}) \pi_{\mathrm{R}}(z_j) p_j^{\mathrm{R}}}{\{1 - \pi_{\mathrm{R}}(z_j)\} p_j^{\mathrm{A}} + \pi_{\mathrm{R}}(z_j) p_j^{\mathrm{R}}} + \sum_{\substack{j \in \mathrm{TT} \\ j \geqslant i}} \exp(\beta z_j) \frac{\exp(\gamma_{\mathrm{I}}) \pi_{\mathrm{I}}(z_j) p_j^{\mathrm{I}} + \exp(\gamma_{\mathrm{T}}) \{1 - \pi_{\mathrm{I}}(z_j)\} p_j^{\mathrm{A}}}{\pi_{\mathrm{I}}(z_j) p_j^{\mathrm{I}} + \{1 - \pi_{\mathrm{I}}(z_j)\} p_j^{\mathrm{A}}} \right] \tag{16}$$

so that Δ_i is the inverse of the expression in square brackets in equation (16) when $\delta_i = 1$. When there is no non-compliance (and no contamination), this reduces to the usual Breslow estimator. However, in general, solving for Δ_i is more difficult than for conventional proportional hazards because the p_j^k depend on other Δ_l, leading to the need for an iterative solution.

However, some simplification can be achieved if we split individuals into parts and use weights. Specifically, we create six 'incarnations' of each individual corresponding to all possible treatment $\times$ class membership groups, and for incarnation c of individual i define

$$w_{ic} = \begin{cases} I_{\{i \in CT\}}, & c = 1, \\ \dfrac{\pi_{\mathrm{I}}(z_i) p_i^{\mathrm{I}}}{\pi_{\mathrm{I}}(z_i) p_i^{\mathrm{I}} + \{1 - \pi_{\mathrm{I}}(z_i)\} p_i^{\mathrm{A}}} I_{\{i \in \mathrm{TT}\}}, & c = 2, \\ \dfrac{\{1 - \pi_{\mathrm{I}}(z_i)\} p_i^{\mathrm{A}}}{\pi_{\mathrm{I}}(z_i) p_i^{\mathrm{I}} + \{1 - \pi_{\mathrm{I}}(z_i)\} p_i^{\mathrm{A}}} I_{\{i \in \mathrm{TT}\}}, & c = 3, \\ \dfrac{\{1 - \pi_{\mathrm{R}}(z_i)\} p_i^{\mathrm{A}}}{\{1 - \pi_{\mathrm{R}}(z_i)\} p_i^{\mathrm{A}} + \pi_{\mathrm{R}}(z_i) p_i^{\mathrm{R}}} I_{\{i \in \mathrm{CC}\}}, & c = 4, \\ \dfrac{\pi_{\mathrm{R}}(z_i) p_i^{\mathrm{R}}}{\{1 - \pi_{\mathrm{R}}(z_i)\} p_i^{\mathrm{A}} + \pi_{\mathrm{R}}(z_i) p_i^{\mathrm{R}}} I_{\{i \in \mathrm{CC}\}}, & c = 5, \\ I_{\{i \in \mathrm{TC}\}}, & c = 6. \end{cases}$$

If, for notational simplicity, we define the 3×6 matrix

$$v = (v_c) = (v_{lc}) = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

somewhat unconventionally (with $l = 1, 2, 3$ corresponding to T, I and R respectively and $c = 1, \ldots, 6$) the score equations can be written as

$$\frac{\delta\{\log(L)\}}{\delta\gamma_l} = \sum_{i,c} v_{lc}\{\delta_i - \exp(\gamma v_c + \beta z_i)\,\Lambda(t_i)\}w_{ic}, \qquad l = \mathrm{T, I, R},$$

$$\frac{\delta\{\log(L)\}}{\delta\beta_l} = \sum_{i,c} z_{li}\{\delta_i - \exp(\gamma v_c + \beta z_i)\,\Lambda(t_i)\}w_{ic}, \qquad l = 1, \ldots, m,$$

$$\Delta_i = \delta_i \left\{ \sum_{\substack{j \geqslant i \\ c}} \exp(\gamma v_c + \beta z_j) w_{jc} \right\}^{-1},$$

where $\gamma = (\gamma_{\mathrm{T}}, \gamma_{\mathrm{I}}, \gamma_{\mathrm{R}})$. We can now eliminate $\Lambda(t_i)$ from the equations to obtain

$$\frac{\delta\{\log(L)\}}{\delta\gamma_l} = \sum_i \delta_i (v_{li}^* - \bar{v}_{li}), \qquad l = \mathrm{T, I, R},$$

$$\frac{\delta\{\log(L)\}}{\delta\beta_l} = \sum_i \delta_i (z_{li} - \bar{z}_{li}), \qquad l = 1, \ldots, m,$$

where

$$v_{li}^* = \sum_c v_{lc} w_{ic},$$

$$\bar{v}_{li} = \sum_{\substack{j \geqslant i \\ c}} v_{lc}\theta_{ljc} \Big/ \sum_{\substack{j \geqslant i \\ c}} \theta_{ljc}$$

and $\theta_{ljc} = w_{jc}\exp(\gamma v_c + \beta z_j)$, and similarly for $\bar{z}_{li}$ except that v_{lc} is replaced by z_{li}. However, estimation is still difficult since the θ_{ljk} still depend on $\Lambda(t_j)$ via the p_{kj}.

Considerable simplification arises by noting that $\Delta_{i-1}^{-1} - \Delta_i^{-1}$ depends only on (γ, β) and Δ_j, $j < i$, so by fixing (γ, β) the maximum likelihood estimation is reduced to a one-dimensional maximization over Δ_1. This leads to a profile likelihood approach to estimating (γ, β).

3.3.3. Method used for maximizing likelihood

Various methods could be used to maximize the full likelihood model. For a smaller number of observations where computational speed is not so much an issue then the Powell method (Press *et al.*, 1992) that does not use derivative information could be used, as this method is reasonably stable. However, for the simulation set we evaluated a hybrid method which involved alternating between maximizing the base-line hazard step sizes and the risk parameters (i.e. the insistor, refuser, treatment and covariate effects). This method was the fastest and worked well on data sets of 500 observations.

For any value of the risk parameters, the simplification that was noted above allows us to find the base-line hazard function that maximizes the likelihood via a one-dimensional maximization. For a fixed base-line hazard it is straightforward to calculate the derivative of the likelihood with respect to the model parameters. The approach that was taken for estimating the risk parameters is based on a quasi-Newton method that uses an approximation to the inverse Hessian: the Broyden–Fletcher–Goldfarb–Shanno method is used to update the Hessian at each iteration (Press *et al.*, 1992). The algorithm starts with initial values for the prognostic parameters and by taking the Hessian to be the unit matrix. We iterate between updating the base-line hazard (for fixed risk parameters) and updating the risk parameters (for fixed base-line hazard) until convergence. To improve the final estimate (after convergence), the procedure is run again, reinitializing the Hessian to the unit matrix. This process does not add much to the running time as the estimate is already close to the 'best' estimate.

4. Simple example and simulations

4.1. Example

To make the calculations more concrete we first consider a simple example in which there are 38 observations split equally into two groups. There are six known insistors, three known refusers, six failures in the control arm and three failures in the treated arm. This is summarized in Table 2 along with the risk sets and intermediate calculations at each failure time. Estimates for the non-iterative Mantel–Haenszel-type estimator can be directly computed from these quantities. The partial and full likelihood estimates require iteration and are summarized in Table 3 along with the non-iterative solutions. Table 2 also displays the partial and full likelihood estimates for the base-line survival function when the other parameters are fixed at their fitted values. This example is for illustration only and substantial variation in the estimates is expected given the small sample size. More realistic simulated data sets are explored in the next section.

4.2. Simulations

We have studied in detail by simulation the performance of five estimators:

Table 2. Example from Section 4.1 with intermediate calculations for the closed form estimate and the estimated base-line survival function from the full and partial likelihoods†

t_i	$\frac{D_i^{CT}}{N_i^{CT}}$	$\frac{D_i^{CC}}{N_i^{CC}}$	$\frac{D_i^{TT}}{N_i^{TT}}$	$\frac{D_i^{TC}}{N_i^{TC}}$	D_i^C	N_i^C	D_i^T	N_i^T	$\frac{D_i^T N_i^T}{N_i^C + N_i^T}$	$\frac{D_i^C N_i^C}{N_i^C + N_i^T}$	*Full likelihood* $\hat{S}_0(t)$	*Partial likelihood* $\hat{S}_0(t)$
0	0/6	0/13	0/16	0/3								
5	0/5	1/10	0/16	0/3	1	7	0	11	0/18	11/18	0.97	0.95
14	1/5	0/8	0/10	0/2	0	6	−1	5	−6/11	0/11	0.93	0.90
16	0/4	1/8	0/10	0/2	1	6	0	6	0/12	6/12	0.89	0.84
21	0/4	0/7	1/9	0/2	0	5	1	5	5/10	0/10	0.85	0.79
24	0/4	1/7	0/8	0/2	1	5	0	4	0/9	4/9	0.81	0.73
33	0/3	1/6	0/7	0/2	1	4	0	4	0/8	4/8	0.77	0.67
43	0/3	0/5	0/6	1/1	−1	4	0	3	0/7	−3/7	0.72	0.61
50	0/1	0/5	1/4	0/0	0	5	1	3	5/8	0/8	0.63	0.53
54	0/1	1/5	0/3	0/0	1	5	0	2	0/7	2/7	0.55	0.45
Sum									0.58	1.91		

†Three patients in group CC and one in CT are censored before the first failure.

Table 3. Values of the five estimators under study for the example in Table 2

Estimator	*Hazard ratio for the following groups:*		
	Treatment	*Insistor*	*Refuser*
Proportional hazards	0.56	NA†	NA†
Mantel–Haenszel	0.30	0.38	0.83
Efficient weights	0.40	—	—
Partial likelihood	0.58	0.53	2.39
Full likelihood	0.34	0.44	1.07

†NA, not applicable.

(a) a classical Cox proportional hazard estimator which ignores non-compliance (PH);
(b) non-iterative approaches suitable for models without covariates using
    (i) standard Mantel–Haenszel weights (MH) or
    (ii) efficient one-step weights (EW);
(c) the partial likelihood estimator (PL);
(d) the estimator based on the full likelihood (FL).

In each case we have also evaluated the variance estimators. For the full likelihood, numerical methods are used to compute the second derivatives of the profile likelihood from which

Table 4. Parameters that were used in the simulations (Section 4.2)

Run	*Probability of insistor/refuser* ($\pi_I = \pi_R$)	*Insistor hazard ratio*	*Refuser hazard ratio*	*Number of covariates*
1	0.00	0.850	1.000	0
2	0.00	0.850	1.000	1
3	0.00	0.850	1.000	2
4	0.05	0.850	1.000	0
5	0.05	0.850	1.000	1
6	0.05	0.850	1.000	2
7	0.05	0.765	1.111	0
8	0.05	0.765	1.111	1
9	0.05	0.765	1.111	2
10	0.05	0.680	1.250	0
11	0.05	0.680	1.250	1
12	0.05	0.680	1.250	2
13	0.10	0.850	1.000	0
14	0.10	0.850	1.000	1
15	0.10	0.850	1.000	2
16	0.10	0.765	1.111	0
17	0.10	0.765	1.111	1
18	0.10	0.765	1.111	2
19	0.10	0.680	1.250	0
20	0.10	0.680	1.250	1
21	0.10	0.680	1.250	2
22	0.20	0.850	1.000	0
23	0.20	0.850	1.000	1
24	0.20	0.850	1.000	2
25	0.20	0.765	1.111	0
26	0.20	0.765	1.111	1
27	0.20	0.765	1.111	2
28	0.20	0.680	1.250	0
29	0.20	0.680	1.250	1
30	0.20	0.680	1.250	2
31	0.10	0.300	1.000	0
32	0.10	0.300	1.000	1
33	0.10	0.300	1.000	2
34	0.20	0.300	1.000	0
35	0.20	0.300	1.000	1
36	0.20	0.300	1.000	2
37	0.10	0.105	1.000	0
38	0.10	0.105	1.000	1
39	0.10	0.105	1.000	2
40	0.20	0.105	1.000	0
41	0.20	0.105	1.000	1
42	0.20	0.105	1.000	2

estimates are calculated. Here we report on estimates of the treatment effect for 200 observations (small sample) and 2000 observations (large sample) equally divided between two treatment groups. The log-hazard for the treatment was either 0.6 for the small sample or 0.85 for the large sample, corresponding to similar standard errors for the treatment effect when there were no insistors or refusers. All test statistics were based on the Wald test for the log-hazard-ratio, which performed slightly better throughout than other standard options. We consider three covariate structures: no covariate, a standard normal covariate independent of group assignment, with coefficient ln(1.2), and a model with two covariates, in which the first was as above but the second was another independent standard normal covariate also with coefficient ln(1.2), but linked to group membership by shifting the mean from 0 to 0.5 in refusers only. 42 different

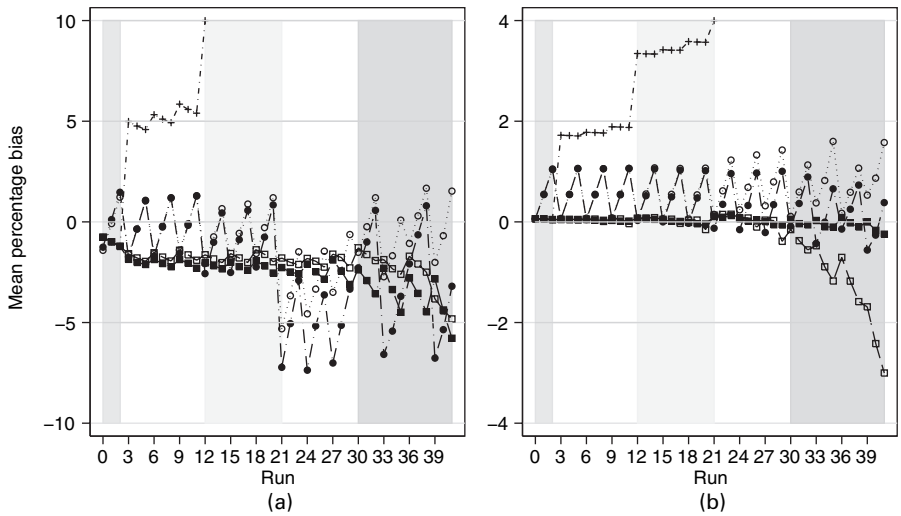

Fig. 2. Mean percentage bias of the estimated treatment effect (■, FL; □, PL; ●, EW; ○, MH; +, PH): (a) $n = 200$; (b) $n = 2000$

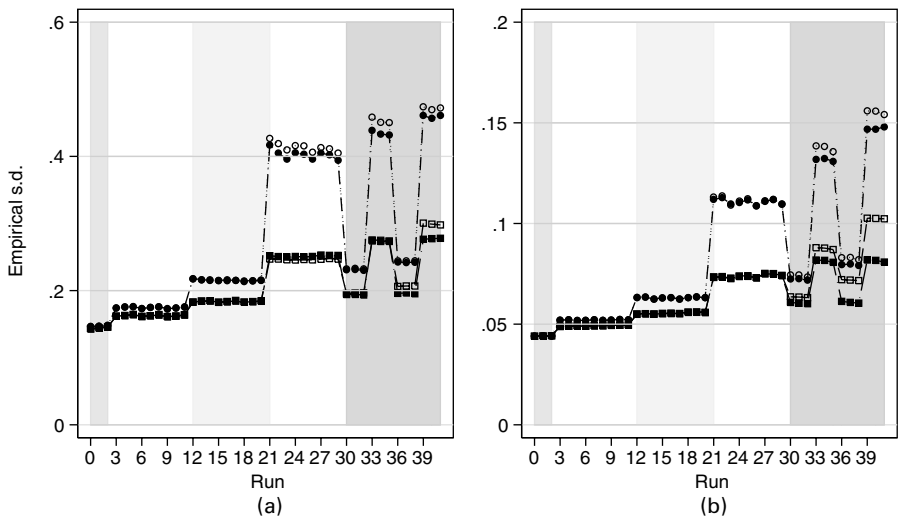

Fig. 3. Empirical standard error of the estimated treatment effect (■, FL; □, PL; ●, EW; ○, MH; +, PH): (a) $n = 200$; (b) $n = 2000$

models with different values of $\pi_I = \pi_R$, γ_R and γ_I and no, one or two covariates were calculated (Table 4). We assume that the insistor and refuser proportions are the same and range from 0 to 20%. Asymmetric models gave similar results. For the first 30 examples insistor and refuser effects were of the same magnitude, but in opposite directions. The last 12 cases looked at more extreme insistor effects without a refuser effect and were asymmetric in this respect.

All simulations are based on a unit exponential base-line failure time (which does not affect the results) and no censoring (which would tend to diminish differences between estimators if introduced). Here we report only the results for the treatment effect. All results are based on 1000 simulations. The 42 different models are generated from the same underlying random

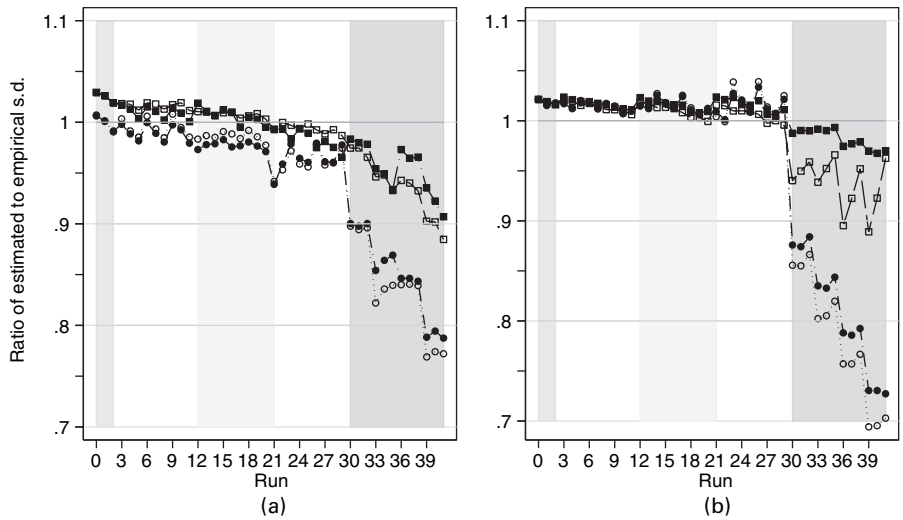

Fig. 4. Ratio of the average estimated standard error to the empirical standard error of the estimated treatment effect (■, FL; □, PL; ●, EW; ○, MH): (a) $n = 200$; (b) $n = 2000$

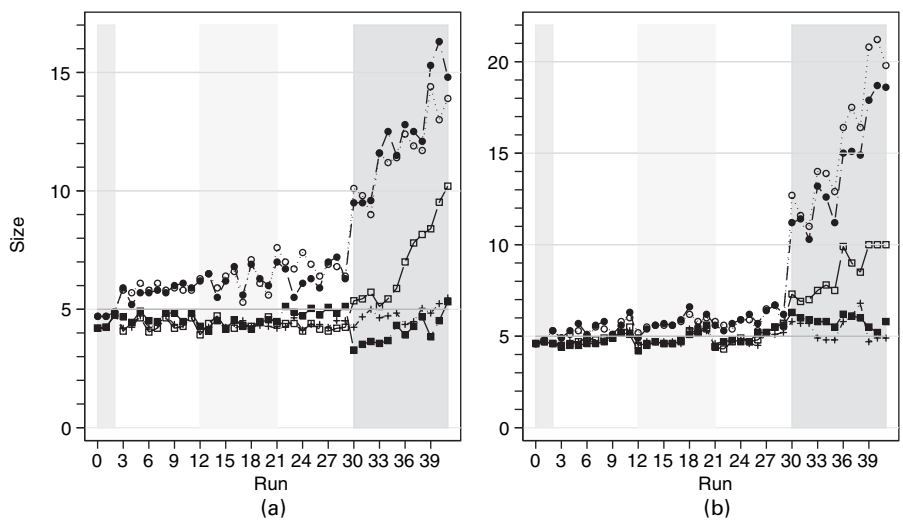

Fig. 5. Size of the Wald test (log-estimate/estimated standard error) when $\ln(\theta_T) = 0$ (■, FL; □, PL; ●, EW; ○, MH; +, PH): (a) $n = 200$; (b) $n = 2000$

variables and so are positively correlated but will give minimum errors for differences between pairs of tests.

Fig. 2(a) shows the mean percentage bias for $n = 200$ and Fig. 2(b) for $n = 2000$. Since the hazard ratio for treatment is less than 1, conservative estimates lead to a positive bias. As expected, underestimation of the treatment effect is apparent for estimator PH (Cox). There is also consistent underestimation for the weighted estimators in models containing covariates (which they ignore), again as expected. In the large sample, estimator PL shows anticonservative bias in the models with large insistor effects, but FL is essentially unbiased for all cases, as is estimator EW when there are no covariates. For the small sample size, both estimators PL and FL are slightly anticonservative, whereas EW and to some extent MH are covariate dependent. The empirical standard errors of θ_T are shown in Fig. 3. Estimators FL and PL performed best and were very similar for cases 1–30. For the remaining more extreme cases FL was better. Estimators MH and EW were substantially worse when there was substantial non-compliance or insistor effects. Details are not shown for estimator PH, as its substantial bias negated any further estimation of precision. A benefit of EW over MH is only likely to appear for larger treatment effects than those considered here.

The ratio of the average predicted to average observed standard error of θ_T is shown in Fig. 4. This was very good for estimators FL and PL for small and large samples, and particularly poor for MH and EW in extreme cases. For extreme cases, an advantage of FL over PL was apparent.

Results for the size of the test are shown in Fig. 5, where the treatment effect is set to 0 ($\theta_T = 1$). All tests are adequate for $n = 2000$ for the first 30 examples, but the MH and EW tests are highly anticonservative for the more difficult cases, reflecting an underestimate of the variance. Estimator PL also has problems in these cases as a result of bias, but the FL and PH methods are adequate for all cases. Similar results are seen for the small sample.

Lastly, in Fig. 6 the power is shown. Estimators FL and PL are similar but, in the difficult cases, FL has slightly larger power. This benefit would be larger if we were to correct the size of estimator PL. The PH estimate gives rise to fairly efficient tests for the first 30 cases, suggesting

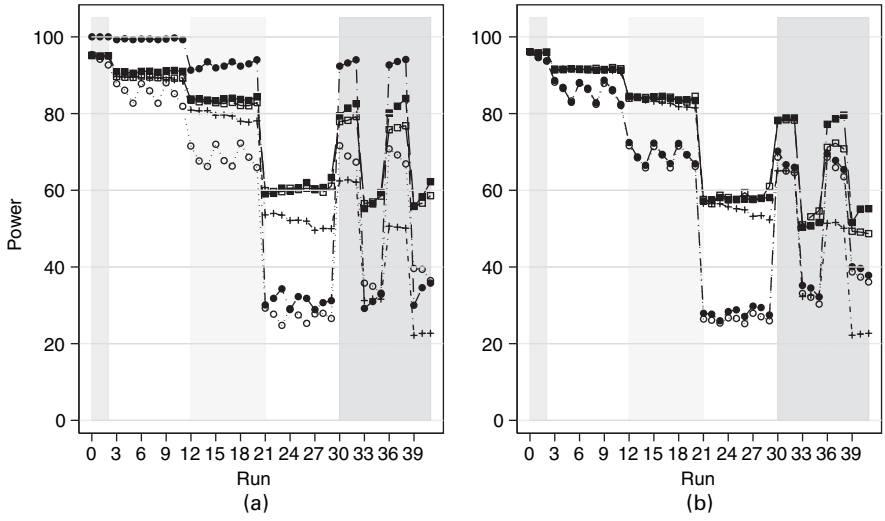

Fig. 6. Power of the Wald test for the treatment effect (■, FL; □, PL; ●, EW; ○, MH; +, PH): (a) $n = 200$; (b) $n = 2000$

that the treatment parameters and their standard error are deflated by similar amounts for the PH case, as suggested previously (Cuzick *et al.*, 1997).

In summary estimator FL performs very well throughout the entire range of cases, for both small and large sample sizes. The PL test is substantially easier to compute and performs well except when there are large insistor (or refuser) effects. Their existence can be checked from the data. Curiously the second covariate that is linked to class membership had little negative effect on the PL, even though this is not accounted for in the model. The EW approach added little to estimator MH: both were substantially less powerful when non-compliance was very large. The PH model provided a powerful test of the null hypothesis in all except extreme cases but gave a strongly biased estimate of the individual treatment effect (which it does not try to measure directly).

The FL method does not require inordinate computing time and details are available from the authors on request.

5. Extensions

Various extensions to the methods are possible.

5.1. Non-compliance during prolonged treatment

Formally it is easy to generalize the model to include non-compliance occurring during prolonged treatment (or repeated screening) by allowing time-dependent treatment or compliance states. However, the interpretation of such a model is more difficult. Treatment changes are often a result of changing disease status, and a wash-out period, in which individuals are kept in their previous treatment group for an interval after any change, is usually prudent. The length of that interval is difficult to judge, but it should be sufficiently long for a change in disease state to become manifest.

5.2. Placebo-controlled trials

It is reasonably straightforward to adjust for placebo-controlled trials. When non-compliance occurs before any treatment is offered, non-compliant patients can simply be excluded as they are equally likely to be in either treatment group and are thus non-informative. Non-compliance after commencement of treatment could be related to disease progression and the *caveats* apply as for 'open label' trials. Non-compliance, after say one cycle of treatment, which is likely to be due to side-effects in the active arm can be managed as above by postulating a similar class in the placebo arm, so long as the drop-out at this point is very infrequent in the placebo arm and the treatment effect after just one cycle can be assumed to be minimal. If there is appreciable non-compliance in the placebo arm as well, further partitioning of the non-compliers into those who are not associated with treatment (determined by the number in the placebo arm) and those who are associated with treatment (determined by subtracting the number in the placebo arm from those in the active arm) is required.

5.3. Choice of a third option

An extension to the choice of a third option is best exemplified by a trial of two active agents, where some patients choose after randomization to take no treatment. If the decision is not based on the allocated treatment, then it is relatively straightforward to include another class for this group. More generally, it becomes difficult to determine the extent to which these individuals are in the same 'class' in the two allocated treatment arms. Choosing a third active treatment also creates difficulties.

5.4. Ties

Ties between censored and uncensored observations are easily handled by assuming that the uncensored observation occurred first. When ties between uncensored observations occur the point estimators are easily modified. For the simple estimator with no covariates, the D_i^σ would be allowed to assume integer values and more than one could be positive at each t_i. For the likelihood-based estimators, a simple approximation is to include more than one term at the relevant failure time, all sharing the same risk set. A more precise approach would downweight the contribution of the tied elements to the risk set, and an exact approach would include terms for all possible permutations of the tied elements. A random splitting of ties is also possible, but the point estimate depends on the random sequence and is inefficient. Averaging several random splittings approximates the full likelihood, but the variance estimate will be too large. Adjustment of the variances is, in general, tedious but the effects are small unless ties are common. Estimates based on the 'simple approximations' or the binomial variances of the D_i^σ are straightforward but will be conservative.

5.5. Randomization

Although it is theoretically possible to use this method with a non-randomized historical or concurrent comparison group, the dangers of non-randomized comparisons outweigh any potential gain in most cases. The method reinforces the value of randomization in ensuring that like is compared with like, but it offers the opportunity to do this specifically in patients who are willing to accept the new treatment. In this way the dangers of non-randomized comparisons are avoided, but we can still obtain unbiased estimates of the true treatment effect at the individual level, as opposed to the population estimate of the effect of offering the treatment in a trial setting, which is what is estimated by conventional intent-to-treat analysis.

Acknowledgements

We thank Stephen Duffy for helpful comments and Rob Edwards for early computing work.

References

Altman, D. G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall.
Bonadonna, G. and Valagussa, P. (1981) Dose-response effect of adjuvant chemotherapy in breast cancer. *New Engl. J. Med.*, **304**, 10–15.
Breslow, N. E. (1974) Covariate analysis of censored survival data. *Biometrics*, **30**, 89–99.
Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.
Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.
Cuzick, J., Edwards, R. and Segnan, N. (1997) Adjusting for non-compliance and contamination in randomised clinical trials. *Statist. Med.*, **16**, 1017–1029.
Duffy, S. W., Cuzick, J., Tabar, L., Vitak, B., Chen, T. H.-H., Yen, M.-F. and Smith, R. A. (2002) Correcting for non-compliance bias in case–control studies to evaluate cancer screening programmes. *Appl. Statist.*, **51**, 235–243.
Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
Loeys, T. and Goetghebeur, E. (2003) A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*, **59**, 100–105.
McKeague, I. W. and Sasieni, P. D. (1994) A partly parametric additive risk model. *Biometrika*, **81**, 501–514.
Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, R. and Smith, P. G. (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I, Introduction and design. *Br. J. Cancer*, **34**, 585–612.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in C: the Art of Scientific Computing*. Cambridge: Cambridge University Press.
Robins, J. M. and Tsiatis, A. A. (1991) Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communs Statist. Theory Meth.*, **20**, 2609–2631.

Shorack, G. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.
Sommer, A. and Zeger, S. L. (1991) On estimating efficacy from clinical trials. *Statist. Med.*, **10**, 45–52.
Zelen, M. (1979) A new design for randomised clinical trials. *New Engl. J. Med.*, **300**, 1242–1245.
Zelen, M. (1990) Randomized consent designs for clinical trials: an update. *Statist. Med.*, **9**, 645–656.