

Statistical Methods in Medical Research

<http://smm.sagepub.com/>

Crossover studies with survival outcomes

Jozefien Buyze and Els Goetghebeur

Stat Methods Med Res published online 29 June 2011

DOI: 10.1177/0962280211402258

The online version of this article can be found at:

<http://smm.sagepub.com/content/early/2011/06/24/0962280211402258>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Statistical Methods in Medical Research* can be found at:

Email Alerts: <http://smm.sagepub.com/cgi/alerts>

Subscriptions: <http://smm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [OnlineFirst Version of Record](#) - Jun 29, 2011

[What is This?](#)

Crossover studies with survival outcomes

Jozefien Buyze and Els Goetghebeur

Statistical Methods in Medical Research
25(0) 1–18

© The Author(s) 2011

Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/0962280211402258

smm.sagepub.com



Abstract

Crossover designs are well known to have major advantages when comparing the effect of two treatments which do not interact. With a right-censored survival endpoint, however, this design is quickly abandoned in favour of the more costly parallel design. Motivated by human immunodeficiency virus (HIV) prevention studies which lacked power, we evaluate what may be gained in this setting and compare parallel with crossover designs. In a heterogeneous population, we find and explain a substantial increase in power for the crossover study using a non-parametric logrank test. With frailties in a proportional hazards model, crossover designs equally lead to substantially smaller variance for the subject-specific hazard ratio (HR), while the population-averaged HR sees negligible gain. Its efficiency benefit is recovered when the population-averaged HR is reconstructed from estimated subject-specific hazard rates. We derive the time point for treatment crossover that optimizes efficiency and end with the analysis of two recent HIV prevention trials. We find that a Cellulose sulphate trial could have hardly gained efficiency from a crossover design, while a Nonoxynol-9 trial stood to gain substantial power. We conclude that there is a role for effective crossover designs in important classes of survival problems.

Keywords

crossover design, frailty models, HIV prevention, microbicides, survival analysis

1 Introduction

Crossover trials are powerful designs when the variation between subjects trumps variation within.¹ They are particularly advantageous for the study of stable chronic diseases and are most straightforward in the absence of a carry-over effect or treatment-by-period interaction when missing data can be avoided.

Perhaps surprisingly, ample advice but relatively little evidence is found to support the use or not of crossover designs for survival outcomes. Senn² and the Cochrane Collaboration³ call them unsuitable for survival analysis unless there are recurrent events, but do not elaborate on the statistical basis for this advice. In a recent publication, Senn⁴ has, however, indicated to have changed his mind. Meanwhile, crossover studies have appeared in the literature occasionally. For example in breast cancer, disease-free survival was compared in a crossover trial with wash-out

Ghent University, Department of Applied Mathematics & Computer Science, Krijgslaan 281-S9, 9000, Gent, Belgium

Corresponding author:

Jozefien Buyze, Ghent University Department of Applied Mathematics & Computer Science, Krijgslaan 281-S9, 9000, Gent, Belgium

Email: Jozefien.Buyze@UGent.be

period;^{5,6} in fertility studies, a marginal proportional odds model has been proposed for analyzing studies with repeated crossover.⁷

Motivated by problems of human immunodeficiency virus (HIV) prevention studies, we examine the potential benefit of crossover trials in survival analysis. Notwithstanding some success stories in HIV prevention (cfr. circumcision^{8,9} and the prevention of mother to child transmission¹⁰), the field of microbicide trials in HIV prevention^{11–17} has seen all too many failed randomized studies with little or no power. This prompted the study on ‘Methodological challenges in biomedical HIV prevention trials’ and the 2008 IOM report by The National Academies Press.¹⁸ The large heterogeneity in risk behaviour between women¹⁹ naturally raises the crossover question in this context. Furthermore, since condoms as microbicide gels require application with every sexual act they seek to protect, their ‘half life’ is quite short and supports a logical argument against carry-over effects. Situations where this can go wrong are discussed in the final section.

We set out to compare parallel and crossover trials of equal total duration or follow-up time in this setting, where initially, we let participants in the crossover design switch treatment ‘halfway’ through the study if they survive the switch point (Figure 1).

An apparent drawback of the survival outcome stems from participants precluded from crossing over upon experiencing an endpoint in the first period. They then add no further information over the second period of the design. Potentially more problematic is the possibility of premature censoring which equally ends the observation period. The crucial assumption must then be made that – given the available information – censoring in itself is not related to subsequent risk. This assumption is often deemed acceptable in standard clinical trials with survival outcome and we will rely on it here too.

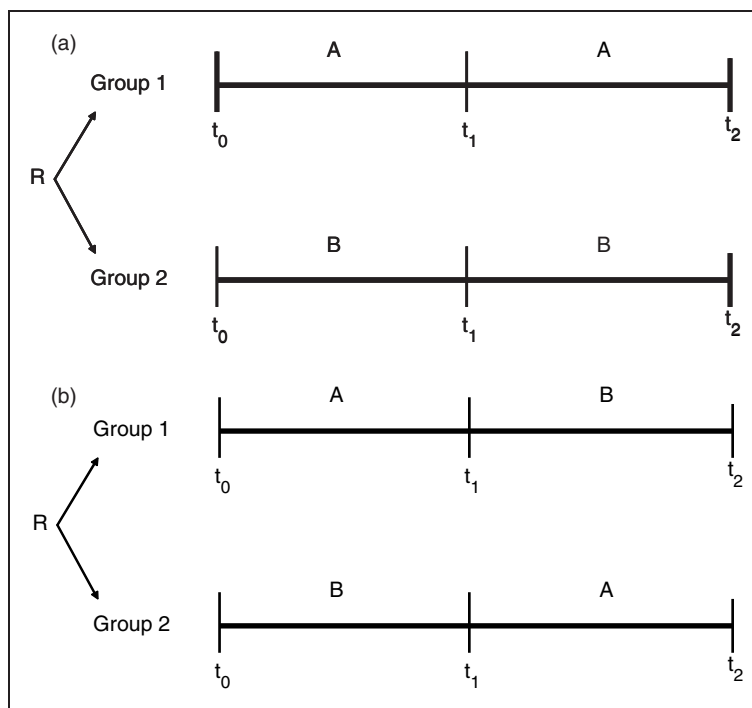


Figure 1. (a) parallel design, (b) crossover design.

In a recent article by Nason and Follmann,²⁰ crossover trials were used with absorbing binary endpoints where similar censoring issues may occur. They, however, ignored this as well as the timing of events within a period of the crossover design. Instead, they applied both a parametric mixture model and a non-parametric test to repeated binary data. Their mixture model assumes further no period effect and may yield a probability of an event above 1 which makes interpretation difficult. In this article, we set out to analyse the actual event times using not only a parametric mixture model and a non-parametric test but also a semi-parametric mixture model. In our approach, baseline hazard may be non-constant over time and the optimal crossover time point is derived.

We detail our setting in Section 2 and compare the power of the (weighted) logrank test in the parallel and crossover designs in Section 3. In Sections 4 and 5, we discuss the benefit of crossover trials for estimation under parametric and semi-parametric working models, respectively. In Section 6, we examine how much could be gained from the crossover design in two recent HIV prevention trials. We derive the change point that optimizes efficiency for the crossover design in Section 7 and end with a discussion and conclusions in Section 8.

2 Setting

When individuals share a common constant baseline hazard, a renewal process is generated: within and between subject comparisons refer to the same (accumulated) hazard. In such setting, the crossover design is expected to only differ from the parallel design in the number of participants that receive treatment A *versus* B in the second period. Simulation confirms that it is slightly more efficient to allocate the largest group of participants at the end of the first period to the higher risk treatment in the second period. The resulting crossover gain in efficiency is negligible, however.

In what follows, we focus on a random sample of individuals $i = 1, \dots, n$ from a heterogeneous population. Let the survival distribution of individual i under control conditions be characterized by an individual-specific hazard at baseline λ_i and the heterogeneity in baseline risks be reflected by frailties γ_i , i.e. $\lambda_i = \lambda \gamma_i$, which can be seen as a reflection of common causes of death over time as in Hernan et al.²¹

Treatment is further assumed to affect individual hazards λ_i^j in treatment group $j = A, B$ proportionally: $\lambda_i^B = \Delta \lambda_i^A$ with Δ the individual-specific hazard ratio (HR). The implied ratio of population-averaged hazards will be time-dependent, i.e. $\bar{\lambda}^B(t) = \Delta^m(t) \bar{\lambda}^A(t)$, with $\bar{\lambda}^B(t) (\bar{\lambda}^A(t))$ the population-averaged hazard at time t under treatment B(A) and $\Delta^m(t)$ a time-varying population-averaged HR.

3 The power of (weighted) logrank tests under both designs

Logrank tests compare hazards of failure over a chosen time window between two treatment groups. Their construction avoids specification of the baseline risks, yielding robustness against such pattern misspecification. Since our population-averaged HR is time-varying, a weighted logrank test²² is suggested as the more powerful version of the test. Below we evaluate the performance of both logrank-type tests and explain obtained results.

Let t_i denote the i th-ordered observed failure time, $R_i(n_i)$ the risk set (size) and d_i the number of events at t_i . Z_i indicates the treatment group of the failed subject at t_i ($Z_i = 0, 1$), and $\bar{Z}_i$ its average among R_i . The weighted logrank statistic

$$T = \sum_{t_i} w_i (Z_i - \bar{Z}_i)$$

with estimated variance V^{22} then has under $H_0 : \Delta = 1$, $\frac{T^2}{V} \sim \chi_1^2$ asymptotically.

Schoenfeld²³ demonstrated how weights w_i proportional to $\ln\left(\frac{\tilde{\lambda}^B(t_i)}{\tilde{\lambda}^A(t_i)}\right)$ achieve maximal power. Ignoring the frailty distribution, one ignores these optimal weights, and could use the decreasing Peto–Prentice weights²⁴ as a simple approximation

$$w_i = \prod_{j \leq i} \left(1 - \frac{d_j}{n_j + 1}\right).$$

We compare empirical powers of the (thus weighted) logrank test between the parallel and the crossover design by simulation. In the simulation, we used exponentially distributed baseline hazards as suggested from the prognostic scores estimated in the COL-1492 study.¹⁴ By way of illustration, data are generated with individual-specific time-constant baseline hazards which are themselves exponentially distributed with a mean of 0.3 per year, approximating the 0.264 per year seen in the high risk population of the Nonoxynol-9 trial,¹² and a sample size chosen to obtain a power ranging between 70% and 80%. A total of 1000 simulated datasets of size 500 with individual-specific HR 1.5 and a total follow-up time of 10 years yielded empirical powers of the standard logrank test for the parallel and crossover design of 68% (95% confidence interval (CI): 65–71%) versus 89% (95% CI: 87–91%), respectively, accompanied by an empirical type I error of 4.7% (95% CI: 3.4–6.0%) and 4.1% (95% CI: 2.9–5.3%), respectively. For the weighted logrank test, this is 73% (95% CI: 70–76%) and 83% (95% CI: 81–85%), respectively, demonstrating a 10% absolute increase in power for the crossover design with now an empirical type I error of 4.7% (95% CI: 3.4–6.0%) in both instances.

The power of the logrank test at the α significance level equals $P(Z \geq z_{1-\alpha/2} - c) + P(Z \leq -z_{1-\alpha/2} - c)$ with $Z \sim N(0, 1)$ and c the non-centrality parameter of the test statistic. Let F be the follow-up time, t_1 the switch point, $\lambda_i^*(s)$ the population-averaged hazard in group i , $X_i(t)$ the expected number of observed survivors at time t in group i and $X(t)$ the overall expected number of observed survivors at time t . Without making any parametric assumptions, the non-centrality parameter c for the parallel design is

$$\frac{\int_0^F \frac{X_0(s)X_1(s)(\Delta^m(s)-1)\lambda_0^*(s)}{X(s)} ds}{\sqrt{\int_0^F \frac{X_0(s)X_1(s)}{X(s)} \lambda_0^*(s) ds + \int_0^F \left(\frac{X_0(s)}{X(s)}\right)^2 X_1(s)(\Delta^m(s)-1)\lambda_0^*(s) ds}} \quad (3.1)$$

and for the crossover design

$$\frac{\int_0^F \frac{X_0(s)X_1(s)(\Delta^m(s)-1)\lambda_0^*(s)}{X(s)} ds}{\sqrt{\int_0^F \frac{X_0(s)X_1(s)}{X(s)} \lambda_0^*(s) ds + \int_0^{t_1} \left(\frac{X_0(s)}{X(s)}\right)^2 X_1(s)(\Delta^m(s)-1)\lambda_0^*(s) ds + \int_{t_1}^F \left(\frac{X_1(s)}{X(s)}\right)^2 X_0(s)(\Delta^m(s)-1)\lambda_0^*(s) ds}} \quad (3.2)$$

The rationale for the non-parametric logrank test gaining power in a crossover design is quite intuitive. Heterogeneity in risk leads to an attenuating population-averaged hazard over time. In the COL-1492 trial¹⁴ for instance, fitted piecewise constant population-averaged hazards decrease from 0.189 in the first 6 months to 0.08 after 3 years.²⁵ In the crossover design, survivors of the first period with the higher risk treatment tend to be less frail than the corresponding survivors on the lower risk treatment. This continues to hold when a third (adjuvant) treatment is added which is effective on both arms (Appendix A). From then on, the less frail group receives the lower risk treatment, increasing the HR further beyond that in the parallel design, hence increasing the power. This rationale holds for a general frailty distribution and a general individual-specific hazard function.

The result stands when loss to follow-up is present (Appendix A). For a conservative sample size correction, one can then multiply the obtained sample size with $\frac{1}{1-l}$ with l the probability of loss to follow-up.

In addition to an increased HR, increased power is owed in the crossover design to an increase in expected number of events. Indeed, the number of events in the two designs coincides up to the switch point with the group of higher risk survivors being smaller than the group of lower risk survivors. In the crossover design, the smaller less frail group next receives the lower risk treatment while the larger more frail group receives the higher risk treatment, yielding more expected events than in the parallel design. In the special case of time-constant individual-specific baseline hazards (Assumption 1), gamma-distributed frailties with mean 1 and variance θ (Assumption 2) and a time-constant individual-specific HR Δ (Assumption 3), expected numbers of events are $\frac{N}{2} \left(2 - \frac{1}{(1+\theta\lambda F)^{1/\theta}} - \frac{1}{(1+\theta\lambda \Delta F)^{1/\theta}} \right)$ in the parallel design and $\frac{N}{2} \left(2 - \frac{1}{(1+\theta\lambda t_1 + \theta\lambda \Delta (F-t_1))^{1/\theta}} - \frac{1}{(1+\theta\lambda \Delta t_1 + \theta\lambda (F-t_1))^{1/\theta}} \right)$ in the crossover design with N the total sample size, F the maximum follow-up time and t_1 the switch point.

Since the population-averaged HR is expected to jump up at the start of the second period before declining again (Figure 2), the Peto–Prentice weights, which are continuously decreasing over time, will not necessarily yield the more efficient weights.

4 Parametric model

Maximum likelihood estimators and corresponding standard errors are easily derived for both the parallel and crossover designs when individual-specific failure times have time-constant hazards with

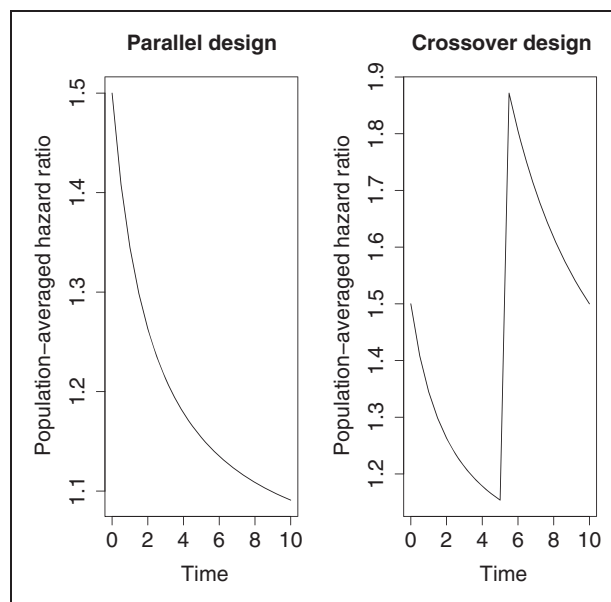


Figure 2. Population-averaged HR over time with exponentially distributed time-constant baseline hazards with mean 0.3 and HR 1.5. Left, parallel design and right, crossover design.

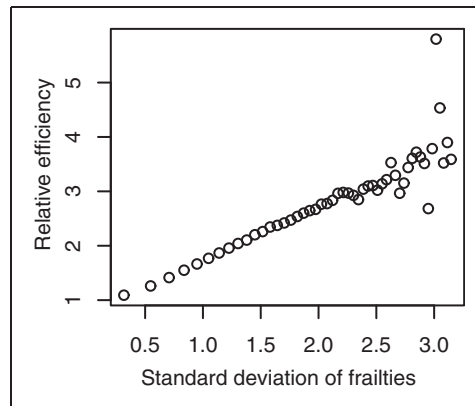


Figure 3. Relative efficiency (crossover design vs. parallel design) with parametric working model (simulated dataset of size 20 000).

Table 1. Simulated datasets of size 2000: results for heterogeneous population in parametric model

	true value	design	empirical mean	mean SE	empirical SE	$\sqrt{\text{MSE}}$	coverage
λ	0.3	Parallel	0.300	0.017	0.017	0.017	0.938
λ	0.3	Crossover	0.300	0.015	0.015	0.015	0.946
Δ	1.5	Parallel	1.506	0.117	0.119	0.119	0.981
Δ	1.5	Crossover	1.502	0.090	0.091	0.091	0.984
$\ln(\Delta)$	0.405	Parallel	0.406	0.078	0.078	0.078	0.987
$\ln(\Delta)$	0.405	Crossover	0.405	0.060	0.061	0.061	0.986

gamma-distributed frailties. Under the parametric assumptions (1–3) of Section 3, solutions to the score equations are written out in Appendix B together with the relative efficiency of the crossover *versus* parallel design. We evaluate efficiency again by simulation, generating 1000 datasets of size 2000, as in Section 3: $\lambda = 0.3$ per year, $\Delta = 1.5$ with a total follow-up time of 10 years (and crossover at year 5). Figure 3 shows how the relative efficiency of the crossover *versus* parallel design (calculated with formulae from Appendix B) increases with a larger standard deviation of the frailties. Results in Table 1 support unbiased point estimates under both designs, but the standard error and root mean squared error (MSE) are substantially reduced under the crossover design. The observed relative efficiency is about 1.69 here (i.e. $(0.117/0.09)^2$ for the HR).

With non-constant hazards, for example if the individual baseline hazards are piecewise constant over time, solutions to the score equations would remain asymptotically unbiased with reduced standard error as well as root MSE under the crossover design.

5 Semi-parametric working model

Now assume unspecified baseline failure times but a time-constant (population-averaged or individual-specific) HR for treatment.

5.1 Parallel design

For analysis, we first consider as a data generating model a true Cox proportional hazards frailty model

$$\lambda_i(t) = \lambda_0(t) \exp(\beta Z_i + B_i)$$

with $Z_i = 0$ (1) for treatment A (B) that accounts for heterogeneity between individuals through the frailty term^{26–28} $\exp(B_i)$ assumed to follow a gamma distribution. Key parameter here is $\exp(\beta)$, the individual-specific treatment HR.

We next develop estimation under a working model which is a standard, i.e. marginal, Cox proportional hazards model:

$$\lambda_i^m(t) = \lambda_0^m(t) \exp(\beta^m Z_i). \quad (5.1)$$

Model (5.1) violates the true data-generating model by assuming a homogeneous population. Its parameter $\exp(\beta^m)$, now represents a time average of the true population-averaged HR.²⁹ If $\exp(B_i)$ is gamma(1,1) distributed and the individual-specific hazards are time-constant, then the true

$$HR_{popavg}(t) = \frac{t + \frac{1}{\lambda}}{t + \frac{1}{\lambda\Delta}}. \quad (5.2)$$

5.2 Crossover design

The data-generating Cox proportional hazards frailty model for analysis is now:

$$\lambda_{ik}(t) = \lambda_0(t) \exp(\beta Z_{ik} + B_i) \quad (5.3)$$

where k ($k=1, 2$) denotes the period, $\exp(B_i)$ is gamma distributed, and the subject-specific treatment HR $\exp(\beta)$ is of key interest.

As a working model, the marginal standard Cox proportional hazards model:

$$\lambda_{ik}^m(t) = \lambda_0^m(t) \exp(\beta^m Z_{ik})$$

is again inconsistent with the true model (5.3) and has $\exp(\beta^m)$ time averaging the true population HR for treatment. If $\exp(B_i)$ is gamma(1,1) distributed and the individual-specific hazards are time-constant, then

$$HR_{popavg}(t) = \begin{cases} \frac{t + \frac{1}{\lambda}}{t + \frac{1}{\lambda\Delta}} & t \leq t_1 \\ \frac{(t-t_1) + \frac{1}{\lambda} + t_1 \Delta}{(t-t_1) + \frac{1}{\lambda\Delta} + \frac{t_1}{\Delta}} & t > t_1 \end{cases}$$

with t_1 the crossing time point.

We compare efficiency of the parallel and crossover design by simulation.

5.3 Simulation results

We generated 1000 datasets of size 2000 as in Section 3. Cox proportional hazards models adapted to the parallel and crossover design as in Section 5.2 estimate a weighted time averaged population

Table 2. Simulated datasets of size 2000: results in semi-parametric model

	time-average HR	Design	True HR	Empirical mean	Mean SE	Empirical SE	$\sqrt{\text{MSE}}$	Coverage
ln(Δ)	Population-averaged	Parallel	0.251	0.248	0.051	0.051	0.051	0.946
ln(Δ)	Population-averaged	Crossover	0.329	0.319	0.051	0.052	0.055	0.937
ln(Δ)	Subject-specific	Parallel	0.405	0.371	0.070	0.122	0.125	0.655
ln(Δ)	Subject-specific	Crossover	0.405	0.397	0.059	0.073	0.074	0.869

HR with weights $S(t)\lambda(t)$ where $S(t)$ is the survival distribution and $\lambda(t)$ the hazard function (see Appendix C for a derivation along the lines of Xu and O'Quigley²⁹):

$$\ln\left(\int_0^{FU} \frac{S(t)\lambda(t)}{\int_0^{FU} S(t)\lambda(t)dt} HR_{popavg}(t)dt\right).$$

It amounts there to 0.251 resp. 0.329 in the parallel and crossover design as shown in Table 2. Under both designs with their respective working models, the estimators perform well on average. For estimation of the subject-specific HR, common to both designs, the crossover design appears two-thirds more efficient ($\text{MSE}(\text{cross}) \leq \frac{1}{3} \text{MSE}(\text{parallel})$). Quite worrying is, however, the coverage, which is unexpectedly low because of an underestimated standard error. In contrast, results are nearly identical when estimating the population-averaged HR. The picture changes when we estimate the population-averaged HR indirectly from the crossover design, by integrating over the individual-specific HR results (with ratio of standard errors then equal to 0.83).

With a different heterogeneity pattern and correct working model, e.g. for normally distributed frailties, we again find smaller bias and standard error for the individual-specific HR in the crossover design. However, when the frailty distribution is then misspecified as gamma, the bias becomes larger and no efficiency benefit remains. Hernan et al.²¹ state indeed that, because frailties represent common causes of failure, not appropriately correcting for an underlying frailty distribution introduces selection bias. On the other hand, the population-averaged HR, though likely biased for the individual-specific effect of treatment on death, represents the causal effect of treatment on the marginal survival distribution, arguably of direct public health interest. In the special case where the frailties of both groups are gamma distributed with the same shape parameter but a different scale parameter, the population-averaged HR is also constant over time³⁰ and the standard Cox proportional hazards model remains correct. We found the crossover design to, then, have a slightly smaller MSE than the parallel design.

6 Relative efficiency in actual trials

Consider two HIV prevention studies, the Cellulose sulphate trial¹³ and the Nonoxynol-9 trial,¹² where microbicides were tested with a negative but non-significant effect. In the Nonoxynol-9 trial,¹² the remaining uncertainty led to the later launch of a large-scale trial with equally unfortunate negative results.¹⁴ We examine whether a crossover study of equal sample size and total duration could have yielded a significant negative effect based on fewer data so that the larger trial might have been avoided or modified. For this purpose, the efficiency of the crossover design is expressed in terms of the expected width of the 95% CI for the population-averaged treatment effect.

6.1 Cellulose sulphate (CONRAD-multisite) trial

The Cellulose sulphate trial¹³ aimed to assess effectiveness of 6% cellulose sulphate gel in preventing male-to-female vaginal transmission of HIV. A total of 1398 women were randomized and followed exactly for 1 year. With an estimated hazard in the placebo group of 0.0333 events per year and population-averaged HR of 1.61 (95% CI [0.86,3.01]), no significant effect of cellulose sulphate gel was found. We simulated data with an individual-specific hazard ratio of 1.63, the value found in a grid search to match a weighted average population HR of 1.61 following equation (5.2). The corresponding 95% CI for the population-averaged HR of a 'simulated crossover result' is expected to be [0.93,2.93], which is hardly smaller and still misses a significant effect for the cellulose sulphate gel. With parametrically generated data, the estimated power to detect a HR of 1.63 through the standard logrank test is 44.2% (95% CI: 39.8–48.6%), respectively 45.2% (95% CI: 40.8–49.6%), in the parallel and crossover design. The empirical power of 500 simulated datasets in function of the sample size is given in Figure 4 and confirms that no substantial gain is to be expected from a crossover design in this setting.

6.2 Nonoxynol-9 trial

One of the early Nonoxynol-9 trials aimed to determine efficacy of the nonoxynol 9 contraceptive sponge in preventing coital acquisition of HIV.¹² A sample size of 116 with 2 years yielded an estimated hazard of 0.264 per year in the placebo group with a HR of 1.7, 95% CI [0.9,3.0]. Following the same approach as above, the individual-specific HR is found to be 1.96 and the 95% CI for the population-averaged HR in the crossover design is expected to be [1.09,3.52]. Hence, the crossover design could have demonstrated a significantly negative effect of the nonoxynol 9 contraceptive sponge and perhaps better informed the later trial. With parametrically generated data, the empirical power of the non-parametric logrank test is 50.4% (95%CI: 46.0–54.8%) respectively 61.6% (95%CI: 57.3–65.9%) in the parallel and crossover design. Figure 4 shows how power is now expected to depend on design as well as sample size (based on 500 simulated datasets).

7 Optimal crossover time point

So far, we assumed participants cross over 'halfway' through the study. We now search for the optimal change point τ_j that maximizes relative efficiency³¹ $\frac{n_0}{n_j}$ with n_0 the required sample size to achieve a given power under the parallel design with specified total duration and n_j the required sample size to achieve that same power for the set change point τ_j and equal total duration. In regular problems, $\frac{n_0}{n_j}$ equals $(\frac{c_j}{c_0})^2$ with c_j the non-centrality parameter of the limiting distribution under contiguous alternatives. For a total follow-up time of 10 years, time-constant individual-specific baseline hazards which are exponentially distributed with mean 0.3 per year and a individual-specific HR of 1.5, the non-centrality parameters can be calculated via Taylor approximations from (3.1) and (3.2). Figure 5 shows relative efficiency of crossover designs *versus* the parallel design in function of the switch point (actual time). Under our assumptions the crossover design is always more efficient and achieves a maximum relative efficiency of 1.8 at a change point of 2 years, which is in this case just one-fifth of the total study length. The right axis in the figure shows the percentage of events which have occurred over time under treatment A (dashed line) and treatment B (dotted line) and reveals how by the optimal crossing time point 50% of the scheduled events for treatment A have occurred.

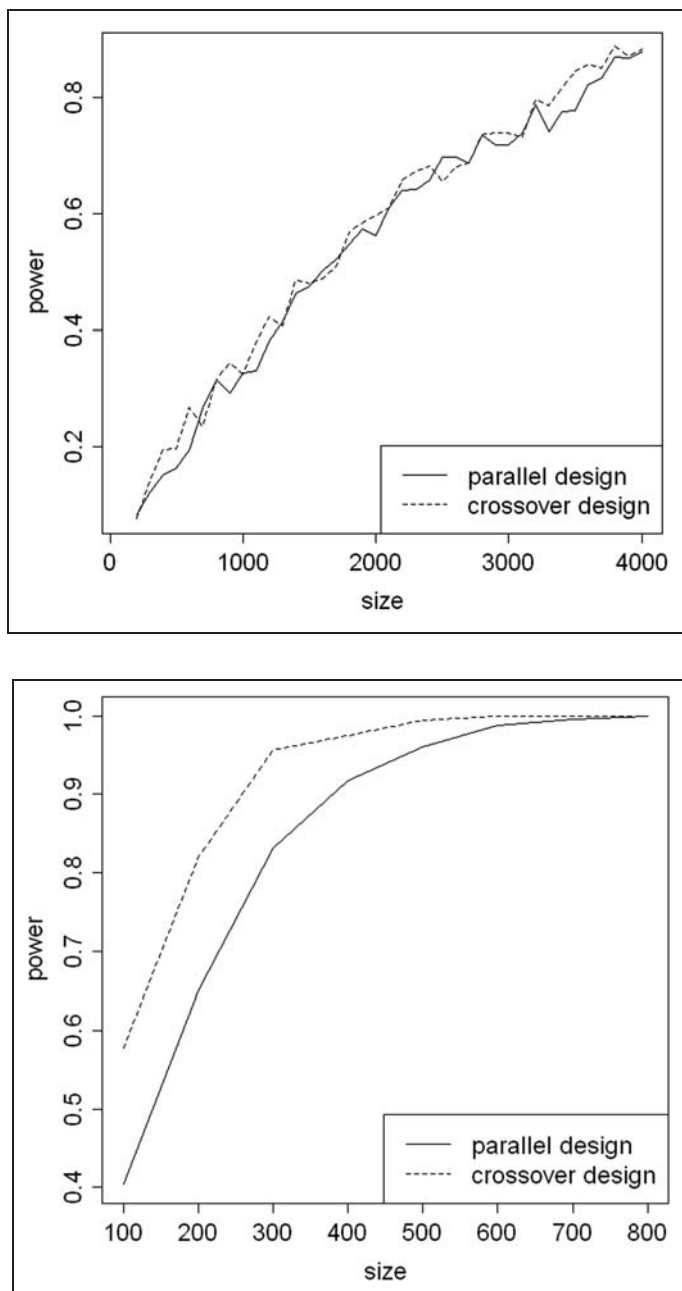


Figure 4. Empirical power (500 simulated datasets) in function of sample size. Top, cellulose sulphate trial and Bottom, nonoxynol-9 trial.

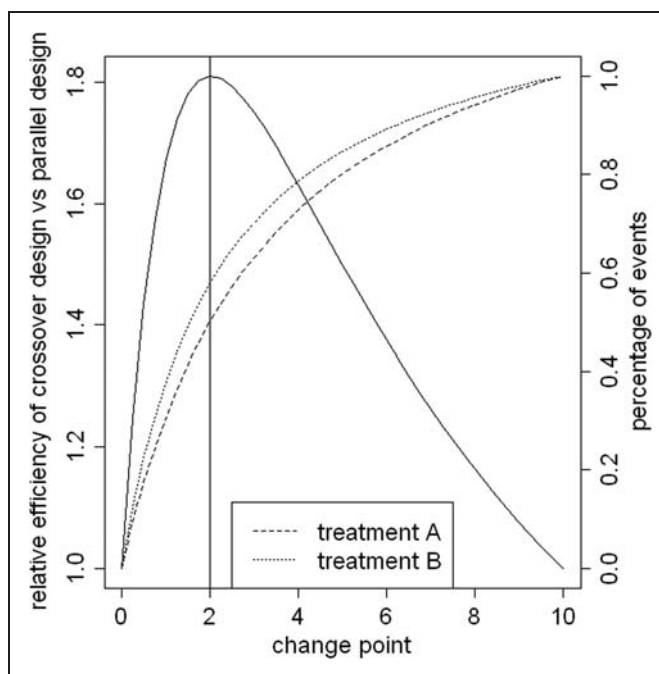


Figure 5. Pitman relative efficiency (crossover design vs. parallel design) vs. change point.

8 Discussion

The literature warns against crossover designs in survival analysis, mostly because individuals who fail in the first period, do not add information over the second period. Subjects who do fail in the second period, however, bring the added information that they survived the first period under the alternative intervention, a fact which may be exploited in the crossover analysis. Our study has revealed that for certain survival time problems, crossover designs are beneficial. Practical barriers and implications of assumptions deserve some discussion. As for continuous and binary outcomes, the obvious gain is achieved under heterogeneity between subjects and absence of carry-over effect. By its very nature, the crossover trial carries limited information on long-term effects and long-term safety following a single treatment. In proof of concept trials, this need not be an issue.

Plausibility of the assumption of no carry-over effect is not all together self-evident in the HIV prevention trials of Section 6. For the N-9 gel, which turned out to be harmful, it has been suggested that the gel causes vaginal lesions which in turn heighten the risk of HIV transmission.^{32,33} Depending on the half life of the lesions, some carry-over effect on the hazard into the next treatment period might well exist. This is however likely of limited impact and if in doubt, a relatively short washout period with condom use only can overcome the problem. On the other hand, when carry-over effects are anticipated, it becomes of intrinsic interest to learn about the AB *versus* BA treatment sequence.⁵ The constant treatment parallel trial lacks information on this. If the switch point is no longer fixed but possibly triggered by an observed event, one enters into the realm of dynamic treatment regimes.^{34,35}

Other standard limitations of crossover designs apply here too. Practical barriers to switching medication may well exist. At the logistic level, a more careful labeling of products for the two

periods is required. At the patient level, a shift in side effects may reveal the nature of the product to the users. For other studies like in epilepsy, changing anti-epileptic drugs is associated with an increased risk of an epilepsy-related event.³⁶ The decision to embark in a crossover *versus* parallel trial will ideally be based on a well-argued balance of all pro's and con's including required sample size and cost.

The simulations shown involved time-constant subject-specific hazards. This, together with the assumed heterogeneity has implications for the shape of the population-averaged hazard distribution. Consistency with the data may be examined through a measure of goodness-of-fit albeit with limited power. Alternatively, one can model the hazard in function of observed characteristics of the participants – such as risk behaviour or preventive measures to learn about heterogeneity. Changing hazards over time influence the optimal change point which occurred in our example when about half of the scheduled events were observed. An attenuating (increasing) population-averaged HR over time tends to shift the optimal change point to the left (right).

The fact that the non-parametric logrank test is generally more powerful in the crossover design is important because it does not require assumptions on the frailty distribution. This benefit can be further exploited for estimation in an accelerated failure time model^{37,38} where backtransformation of observation times are followed by the logrank test. The 95% CI for the treatment effect then consists of these effects for which the test does not reject equality of the backtransformed times at the 5% significance level. The higher power of the logrank test in the crossover design will then yield a shorter CI.

In the presence of heterogeneity in risk, conducting a crossover design is one approach to increase power. Alternatively, one may gain power by explaining part of the frailty variance by covariates which measure baseline risk behaviour (e.g. number of sex partners in HIV prevention studies). In clinical trials, robustness is highly treasured for the intention-to-treat result. Giving it up encounters several specific hurdles here. Measurement of such covariates is complicated due to recall bias, stigma and the private nature of the requested data. Even if correctly measured, they tend to be high dimensional complicating estimation of an adjusted baseline hazard. Their prognostic value may be reduced for the new trial since risk behaviour may well shift during the study as participants get paid, receive medical care and education on risk behaviour. Hence explaining a proportion of the frailty variance by correcting for covariates is not always a feasible or preferred option in practice.

We conclude that, also in survival analysis, the crossover design may bring added efficiency and insights which deserve to be explored further.

References

1. Jones B and Kenward MG. *Design and analysis of cross-over trials*. London: Chapman and Hall, 1989.
2. Senn S. Cross-over trials in Statistics in Medicine: the first 25 years. *Stat Med* 2006; **25**: 3430–3442.
3. Cochrane Collaboration. 2007; <http://www.cochrane-net.org/openlearning/html/modA2-3.htm> (accessed 5 January 2011).
4. Senn S. Cross-over design. In: Chow S (ed.) *Encyclopedia of biopharmaceutical statistics*. London: Informa Healthcare, 2010, pp.380–385.
5. The BIG 1-98 Collaborative Group. Letrozole therapy alone or in sequence with Tamoxifen in women with breast cancer. *N Engl J Med* 2009; **361**: 766–776.
6. Bryant J and Wolmark N. Letrozole after Tamoxifen for breast cancer - What is the price of success? *N Eng J Med* 2003; **349**: 1855–1857.
7. Cohlen BJ, te Velde ER, van Kooij RJ, Looman CWN and Habbema JDF. Controlled ovarian hyperstimulation and intrauterine insemination for treating male subfertility: a controlled study. *Hum Reprod* 1998; **13**: 1553–1558.
8. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R and Puren A. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. *PLoS Med* 2005; **2**: e298.
9. Bailey RC, Moses S, Parker CB, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 2007; **369**: 643–656.

10. Birkhead GS, Pulver WP, Warren BL, et al. Progress in prevention of mother-to-child transmission of HIV in New York State: 1988-2008. *J Public Health Manag Pract* 2010; **16**: 481-491.
11. Lagakos SW and Gable AR. Challenges to HIV prevention - Seeking effective measures in the absence of a vaccine. *N Eng J Med* 2008; **358**: 1543-1545.
12. Kreiss J, Ngugi E, Holmes K, et al. Efficacy of nonoxynol 9 contraceptive sponge use in preventing heterosexual acquisition of HIV in Nairobi prostitutes. *J Am Med Assoc* 1992; **268**: 477-482.
13. Van Damme L, Govinden R, Mirembe FM, et al. Lack of effectiveness of cellulose sulfate gel for the prevention of vaginal HIV transmission. *N Eng J Med* 2008; **359**: 463-472.
14. Van Damme L, Ramjee G, Alary M, et al. Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomised controlled trial. *Lancet* 2002; **360**: 971-977.
15. Feldblum PJ, Adeiga A, Bakare R, et al. SAVVY vaginal gel (C31G) for prevention of HIV infection: a randomized, controlled trial in Nigeria. *PLoS Clin Trials* 2008; **3**: e1474.
16. Stein ZA, Myer L and Susser M. The design of prophylactic trials for HIV: the case of microbicides. *Epidemiology* 2003; **14**: 80-83.
17. Minnis AM and Padian NS. Effectiveness of female controlled barrier methods in preventing sexually transmitted infections and HIV: current evidence and future research directions. *Sex Transm Infect* 2005; **81**: 193-200.
18. Institute of Medicine. *Methodological challenges in biomedical HIV prevention trials*. Washington, DC: The National Academies Press, 2008.
19. Smit J, Middelkoop K, Myer L, et al. Sexual risk factors associated with volunteering for HIV vaccine research in South Africa. *Aids Care* 2006; **18**: 569-573.
20. Nason M and Follmann D. Design and analysis of crossover trials for absorbing binary endpoints. *Biometrics* 2010; **66**: 958-965.
21. Hernan MA, Hernandez-Diaz S and Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615-625.
22. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*. New Jersey: John Wiley & Sons, 2002.
23. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; **68**: 316-319.
24. Peto R and Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Series A* 1972; **135**: 185-207.
25. Bontinck E. The impact of a mixture of hazards on the power of an HIV prevention study. Unpublished master thesis, 2008.
26. Therneau TM, Grambsch PM and Pankratz VS. Penalized survival models and frailty. *J Computat Graph Stat* 2003; **12**: 156-175.
27. Duchateau L and Janssen P. *The frailty model*. New York: Springer, 2008.
28. Gua Go M, Sun L and Huang C. A universal procedure for parametric frailty models. *J Stat Comput Simul* 2004; **74**: 1-13.
29. Xu R and O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics* 2000; **1**: 423-439.
30. Aalen O and Hjort NL. Frailty models that yield proportional hazards. *Stat Probab Lett* 2002; **58**: 335-342.
31. Pitman E. *Some basic theory of statistical inference*. London: Chapman and Hall, 1979.
32. Iyer V and Poddar SS. Update on nonoxynol-9 as vaginal spermicide. *Eur J Contracept Reprod Health Care* 2008; **13**: 339-350.
33. Niruthisard S, Roddy RE and Chutivongse S. The effects of frequent nonoxynol-9 use on the vaginal and cervical mucosa. *Sexually Transmitted Diseases* 18: 176-179.
34. Lavori PW and Dawson R. Adaptive treatment strategies in chronic disease. *Annu Rev Med* 2008; **59**: 443-453.
35. Murphy SA, van der Laan MJ and Robins JM. Marginal mean models for dynamic regimes. *J Am Stat Assoc* 2001; **96**: 1410-1423. (Conduct Problems Prevention Research Group).
36. Hansen RN, Campbell JD and Sullivan SD. Association between antiepileptic drug switching and epilepsy-related events. *Epilepsy Behav* 2009; **15**: 481-485.
37. Robins JM, Blevins D, Ritter G and Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**: 319-336.
38. Branson M and Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Stat Med* 2002; **21**: 2449-2463.
39. Marshall AW and Olkin I. *Life distributions: structure of nonparametric, semiparametric and parametric families*. New York: Springer, 2007, pp. 52-58, 310.

Appendix A

Survivors at t_1 with high risk treatment over $[t_0, t_1]$ are less frail than t_1 -survivors with low risk treatment over $[t_0, t_1]$

In this section, we show that the survivors following high-risk treatment are less frail than the survivors following low-risk treatment. First, we leave the frailty distribution and hazard function unspecified; second, we take into account loss to follow-up.

A.1 General hazard distribution - no loss of follow-up

For a general frailty distribution and individual-specific hazard function, let U_h denote the frailties of the t_1 -survivors following high-risk treatment and U_l the frailties of the t_1 -survivors following

low-risk treatment. The densities, are respectively, f and g where

$$f(u) = \frac{f_0(u)S_0(t_1|u)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)d\gamma_i}$$

and

$$g(u) = \frac{f_0(u)S_1(t_1|u)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)d\gamma_i}$$

Following developments in Marshall and Olkin,³⁹ we first show that U_h is smaller than U_l in the likelihood ratio ordering, i.e. we show that

$$f(u)g(v) \geq f(v)g(u) \quad \forall u \leq v.$$

³⁹ Proof:

$$\begin{aligned}
 & u \leq v \\
 & \Downarrow \\
 & (1 - \Delta)u \leq (1 - \Delta)v \\
 & \Downarrow \\
 & u + \Delta v \leq v + \Delta u \\
 & \Downarrow \\
 & u \int_0^{t_1} \lambda(t)dt + \Delta v \int_0^{t_1} \lambda(t)dt \leq v \int_0^{t_1} \lambda(t)dt + \Delta u \int_0^{t_1} \lambda(t)dt \\
 & \Downarrow \\
 & \exp\left(-\int_0^{t_1} \lambda(t)udt\right) \exp\left(-\int_0^{t_1} \lambda(t)\Delta vdt\right) \geq \exp\left(-\int_0^{t_1} \lambda(t)vdt\right) \exp\left(-\int_0^{t_1} \lambda(t)u\Delta dt\right) \\
 & \Downarrow \\
 & S_0(t_1|u)S_1(t_1|v) \geq S_0(t_1|v)S_1(t_1|u) \\
 & \Downarrow \\
 & \frac{f_0(u)S_0(t_1|u)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)d\gamma_i} \frac{f_0(v)S_1(t_1|v)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)d\gamma_i} \geq \frac{f_0(v)S_0(t_1|v)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)d\gamma_i} \frac{f_0(u)S_1(t_1|u)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)d\gamma_i} \\
 & \Downarrow \\
 & f(u)g(v) \geq f(v)g(u)
 \end{aligned} \tag{A.1}$$

Thus, we have shown that $U_h <_{lr} U_l$ which implies that $U_h <_{hr} U_l$ ³⁹ and from this follows that U_h is stochastically smaller than U_l .³⁹ In the second period of the crossover design, the less frail group (high risk survivors) receives the lower risk treatment, increasing the HR further beyond that in the parallel design. In addition, the number of high-risk survivors is lower than the number of low-risk survivors. Thus, in the second period, the larger group is assigned to the higher risk treatment yielding more events. From both arguments follows that the power is higher in the crossover design than in the parallel design.

A.2 Allowing for loss of follow-up, third treatment

Assume that some observations are lost to follow-up. Let the survival function of duration of follow-up D be denoted by $S_h(t)$ in the high-risk treatment group and $S_l(t)$ in the lower risk treatment group. Then, the densities of U_h and U_l are respectively

$$f(u) = \frac{f_0(u)S_0(t_1|u)S_h(t_1)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)S_h(t_1)d\gamma_i} \quad (\text{A.2})$$

and

$$g(u) = \frac{f_0(u)S_1(t_1|u)S_l(t_1)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)S_l(t_1)d\gamma_i} \quad (\text{A.3})$$

Assuming the loss to follow-up is independent of the survival process, the derivation above remains unaltered until (A.1) and continues as follows:

$$\begin{aligned} S_0(t_1|u)S_1(t_1|v) &\geq S_0(t_1|v)S_1(t_1|u) \\ &\Downarrow \\ &\left\{ \frac{f_0(u)S_0(t_1|u)S_l(t_1)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)S_l(t_1)d\gamma_i} \frac{f_0(v)S_1(t_1|v)S_h(t_1)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)S_h(t_1)d\gamma_i} \right. \\ &\geq \left. \frac{f_0(v)S_0(t_1|v)S_l(t_1)}{\int_0^\infty f_0(\gamma_i)S_0(t_1|\gamma_i)S_l(t_1)d\gamma_i} \frac{f_0(u)S_1(t_1|u)S_h(t_1)}{\int_0^\infty f_0(\gamma_i)S_1(t_1|\gamma_i)S_h(t_1)d\gamma_i} \right\} \\ &\Downarrow \\ f(u)g(v) &\geq f(v)g(u) \end{aligned}$$

We have therefore again shown that $U_h <_{lr} U_l$ and thus, U_h is stochastically smaller than U_l . Hence, the hazard ratio in the crossover design increases compared to the parallel design. Even if there is loss to follow-up in the first period, the expected number of events until t_1 is the same in the parallel and crossover design. In the second period, if there is considerable more loss to follow-up with the high-risk treatment (e.g. because of toxicity), there may be less events in the crossover design than in the parallel design and this might counterbalance the higher HR. However, this will rarely occur in practice.

Now assume that there is another treatment (e.g. risk behaviour counseling) which is more effective early on in the study. The densities of U_h and U_l are of the same form as above (A.2 and A.3) and the same proof holds.

In the specific case that the frailties are gamma distributed with mean 1 and variance θ and a time-constant hazard, $U_h \sim \text{Gamma}\left(\frac{1}{\theta}, \frac{\theta}{\theta\lambda t_1 + 1}\right)$ and $U_l \sim \text{Gamma}\left(\frac{1}{\theta}, \frac{\theta}{\theta\lambda\Delta t_1 + 1}\right)$. A gamma distribution with parameters ν and $\frac{1}{\lambda}$ is the distribution of waiting times for the ν th jump in the Poisson process with parameter λ .³⁹ In the high-risk group the scale parameter $\frac{1}{\lambda}$ of the gamma distribution is smaller than in the low-risk group and thus models the density of smaller waiting times. Hence, $P_h(\gamma > x) \leq P_l(\gamma > x)$ and the gamma distribution in the high-risk group is stochastically smaller than in the low-risk group. We conclude that the population of survivors following high-risk treatment are less frail than the corresponding population survivors following low-risk treatment.

Appendix B

Estimating equations under parametric assumptions

When the frailties are gamma distributed with mean 1 and variance θ , score equations for the parallel design are

$$\begin{cases} \sum_{i=1}^n \delta_{1,i} R_i \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta t_i}{1 + \theta \lambda t_i} \right) + \delta_{1,i} (1 - R_i) \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \Delta t_i}{1 + \theta \lambda \Delta t_i} \right) \\ + (1 - \delta_{1,i}) R_i \left(-\frac{1}{\theta} \frac{\theta t_i}{1 + \theta \lambda t_i} \right) + (1 - \delta_{1,i}) (1 - R_i) \left(-\frac{1}{\theta} \frac{\theta \Delta t_i}{1 + \theta \lambda \Delta t_i} \right) = 0 \\ \sum_{i=1}^n \delta_{1,i} (1 - R_i) \left(\frac{1}{\Delta} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \lambda t_i}{1 + \theta \lambda \Delta t_i} \right) + (1 - \delta_{1,i}) (1 - R_i) \left(-\frac{1}{\theta} \frac{\theta \lambda t_i}{1 + \theta \lambda \Delta t_i} \right) = 0 \end{cases}$$

In the crossover design, the estimating equations are

$$\begin{cases} \sum_{i=1}^n R_i \delta_{1,i} \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta t_i}{1 + \theta \lambda t_i} \right) + (1 - R_i) \delta_{1,i} \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \Delta t_i}{1 + \theta \lambda \Delta t_i} \right) \\ + R_i (1 - \delta_{1,i}) \delta_{2,i} \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta t_1 + \theta \Delta(t_i - t_1)}{1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1)} \right) + (1 - R_i) (1 - \delta_{1,i}) \delta_{2,i} \left(\frac{1}{\lambda} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \Delta t_1 + \theta(t_i - t_1)}{1 + \theta \lambda \Delta t_1 + \theta \lambda(t_i - t_1)} \right) \\ + R_i (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta t_1 + \theta \Delta(t_i - t_1)}{1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1)} \right) + (1 - R_i) (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta \Delta t_1 + \theta(t_i - t_1)}{1 + \theta \lambda \Delta t_1 + \theta \lambda(t_i - t_1)} \right) = 0 \\ \sum_{i=1}^n (1 - R_i) \delta_{1,i} \left(\frac{1}{\Delta} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \lambda t_i}{1 + \theta \lambda \Delta t_i} \right) + R_i (1 - \delta_{1,i}) \delta_{2,i} \left(\frac{1}{\Delta} - \left(\frac{1}{\theta} + 1 \right) \frac{\theta \lambda(t_i - t_1)}{1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1)} \right) \\ + (1 - R_i) (1 - \delta_{1,i}) \delta_{2,i} \left(-\frac{1}{\theta} \frac{\theta \lambda t_1}{1 + \theta \lambda \Delta t_1 + \theta \lambda \Delta(t_i - t_1)} \right) + R_i (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta \lambda(t_2 - t_1)}{1 + \theta \lambda t_1 + \theta \lambda \Delta(t_2 - t_1)} \right) \\ + (1 - R_i) (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta \lambda t_1}{1 + \theta \lambda \Delta t_1 + \theta \lambda \Delta(t_2 - t_1)} \right) = 0 \end{cases}$$

We solve these equations by Newton–Raphson where we take the estimates from an assumed homogeneous population as starting values.

The negative of the information matrix in the parallel design exists of

$$\begin{aligned} \frac{\partial^2 \ln(L(\lambda))}{\partial \lambda^2} &= \sum_{i=1}^n \delta_{1,i} R_i \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 t_i^2}{(1 + \theta \lambda t_i)^2} \right) \\ &\quad + \delta_{1,i} (1 - R_i) \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \Delta^2 t_i^2}{(1 + \theta \lambda \Delta t_i)^2} \right) \\ &\quad + (1 - \delta_{1,i}) R_i \left(\frac{1}{\theta} \frac{\theta^2 t_1^2}{(1 + \theta \lambda t_1)^2} \right) + (1 - \delta_{1,i}) (1 - R_i) \left(\frac{1}{\theta} + \frac{\theta^2 \Delta^2 t_1^2}{(1 + \theta \lambda \Delta t_1)^2} \right) \\ \frac{\partial^2 \ln(L(\lambda))}{\partial \Delta^2} &= \sum_{i=1}^n \delta_{1,i} (1 - R_i) \left(-\frac{1}{\Delta^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \lambda^2 t_i^2}{(1 + \theta \lambda \Delta t_i)^2} \right) \\ &\quad + (1 - \delta_{1,i}) (1 - R_i) \left(\frac{1}{\theta} \frac{\theta^2 \lambda^2 t_1^2}{(1 + \theta \lambda \Delta t_1)^2} \right) \\ \frac{\partial^2 \ln(L(\lambda))}{\partial \lambda \partial \Delta} &= \sum_{i=1}^n \delta_{1,i} (1 - R_i) \left(-\left(\frac{1}{\theta} + 1 \right) \frac{\theta t_i}{(1 + \theta \lambda \Delta t_i)^2} \right) \\ &\quad + (1 - \delta_{1,i}) (1 - R_i) \left(-\frac{t_1}{(1 + \theta \lambda \Delta t_1)^2} \right) \end{aligned}$$

and in the crossover design

$$\begin{aligned}
\frac{\partial^2 \ln(L(\lambda))}{\partial \lambda^2} &= \sum_{i=1}^n R_i \delta_{1,i} \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 t_i^2}{(1 + \theta \lambda t_i)^2} \right) \\
&\quad + (1 - R_i) \delta_{1,i} \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \Delta^2 t_i^2}{(1 + \theta \lambda \Delta t_i)^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) \delta_{2,i} \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{(\theta t_1 + \theta \Delta(t_i - t_1))^2}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) \delta_{2,i} \left(-\frac{1}{\lambda^2} + \left(\frac{1}{\theta} + 1 \right) \frac{(\theta \Delta t_1 + \theta(t_i - t_1))^2}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_i - t_1))^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(\frac{1}{\theta} \frac{(\theta t_1 + \theta \Delta(t_2 - t_1))^2}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_2 - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(\frac{1}{\theta} \frac{(\theta \Delta t_1 + \theta(t_2 - t_1))^2}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_2 - t_1))^2} \right) \\
\frac{\partial^2 \ln(L(\lambda))}{\partial \Delta^2} &= \sum_{i=1}^n (1 - R_i) \delta_{1,i} \left(-\frac{1}{\Delta^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \lambda^2 t_i^2}{(1 + \theta \lambda \Delta t_i)^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) \delta_{2,i} \left(-\frac{1}{\Delta^2} + \left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \lambda^2 (t_i - t_1)^2}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) \delta_{2,i} \left(\left(\frac{1}{\theta} + 1 \right) \frac{\theta^2 \lambda^2 t_1^2}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_i - t_1))^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(\frac{1}{\theta} \frac{\theta^2 \lambda^2 (t_2 - t_1)^2}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_2 - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(\frac{1}{\theta} \frac{\theta^2 \lambda^2 t_1^2}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_2 - t_1))^2} \right) \\
\frac{\partial^2 \ln(L(\lambda))}{\partial \lambda \partial \Delta} &= \sum_{i=1}^n (1 - R_i) \delta_{1,i} \left(-\left(\frac{1}{\theta} + 1 \right) \frac{\theta t_i}{(1 + \theta \lambda \Delta t_i)^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) \delta_{2,i} \left(-\left(\frac{1}{\theta} + 1 \right) \frac{\theta(t_i - t_1)}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_i - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) \delta_{2,i} \left(-\left(\frac{1}{\theta} + 1 \right) \frac{\theta t_1}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_i - t_1))^2} \right) \\
&\quad + R_i (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta(t_2 - t_1)}{(1 + \theta \lambda t_1 + \theta \lambda \Delta(t_2 - t_1))^2} \right) \\
&\quad + (1 - R_i) (1 - \delta_{1,i}) (1 - \delta_{2,i}) \left(-\frac{1}{\theta} \frac{\theta t_1}{(1 + \theta \lambda \Delta t_1 + \theta \lambda(t_2 - t_1))^2} \right)
\end{aligned}$$

The variance-covariance matrix of the estimators is estimated as the inverse of the information matrix.

Appendix C

Weighted average population HR over time

Suppose the true model is

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\}$$

and the working model for analysis is the standard Cox proportional hazards model

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta Z(t)\}.$$

Xu and O'Quigley²⁹ state that the maximum partial likelihood estimator converges to the solution of the equation

$$\int_0^\infty \left\{ \frac{s^{(1)}(\beta(t), t)}{s^{(0)}(\beta(t), t)} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} s^{(0)}(\beta(t), t) \lambda_0(t) dt = 0 \quad (\text{C.1})$$

with

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta Z_i(t)} Z_i(t)^r, \quad s^{(r)}(\beta, t) = E[S^{(r)}(\beta, t)]$$

and

$$S^{(r)}(\beta(t), t) = n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta(t) Z_i(t)} Z_i(t)^r, \quad s^{(r)}(\beta(t), t) = E[S^{(r)}(\beta(t), t)]$$

where the expectations are taken with respect to the true distribution and $Y(t)$ at the risk indicator.

Then, analogous as in Section 3 of Xu and O'Quigley,²⁹ we denote the solution of (C.1) by β^* and write the estimating equation (C.1) as

$$\int_0^\infty (\beta(t) - \beta^*) v(\beta, t) s^{(0)}(\beta(t), t) \lambda_0(t) dt \approx 0$$

$$\beta^* \approx \frac{\int_0^\infty \beta(t) v(\beta, t) s^{(0)}(\beta(t), t) \lambda_0(t) dt}{\int_0^\infty v(\beta, t) s^{(0)}(\beta(t), t) \lambda_0(t) dt}$$

Assuming that $v(\beta, t)$ changes little with time, we find

$$\beta^* \approx \frac{\int_0^\infty \beta(t) s^{(0)}(\beta(t), t) \lambda_0(t) dt}{\int_0^\infty s^{(0)}(\beta(t), t) \lambda_0(t) dt}$$

and

$$\begin{aligned} s^{(0)}(\beta(t), t) \lambda_0(t) &= E \left[n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta(t) Z_i(t)} \lambda_0(t) \right] \\ &= E \left[n^{-1} \sum_{i=1}^n Y_i(t) \lambda(t, Z_i(t)) \right] \\ &\approx S(t) S_C(t) \lambda(t, Z(t)) \end{aligned}$$

so

$$\beta^* \approx \frac{\int_0^\infty S(t) S_C(t) \lambda(t, Z(t)) \beta(t) dt}{\int_0^\infty S(t) S_C(t) \lambda(t, Z(t)) dt}$$