How serious is bias in effect estimation in randomised trials with survival data given risk heterogeneity and informative censoring?

DOI:
[10.1002/sim.7343](https://doi.org/10.1002/sim.7343)

Document Version
Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

McNamee, R. (2017). How serious is bias in effect estimation in randomised trials with survival data given risk heterogeneity and informative censoring? *Statistics in medicine*. <https://doi.org/10.1002/sim.7343>

Published in:
Statistics in medicine

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Title: How serious is bias in effect estimation in randomised trials with survival data given risk heterogeneity and informative censoring?

Author and (corresponding author): Roseanne McNamee, Professor

Affiliation and contact address: Centre for Biostatistics
University of Manchester
Oxford Road
Manchester M13 9PL, UK

Abstract

It is often assumed that randomisation will prevent bias in estimation of treatment effects from clinical trials but this is not true of the semiparametric Proportional Hazards model for survival data when there is underlying risk heterogeneity. Here, a new formula is proposed for estimation of this bias, improving on a previous formula through ease of use and clarity regarding the role of the mid-study cumulative hazard rate, shown to be an important factor for the bias magnitude.

Informative censoring (IC) is recognised as a source of bias. Here, work on selection effects among survivors due to risk heterogeneity is extended to include IC. A new formula shows that bias in causal effect estimation under IC has two sources; one consequent on heterogeneity and one from the additional impact of IC. The formula provides new insights not previously shown: there may be less bias under IC than when there is no IC and even, in principle, zero bias.

When tested against simulated data, the new formulae are shown to be very accurate for prediction of bias in Proportional Hazards and Accelerated Failure Time (AFT) analyses which ignore heterogeneity. These data are also used to investigate the performance of AFT models which explicitly model risk heterogeneity ('frailty models') and G estimation. These methods remove bias when there is simple censoring but not with informative censoring when they may lead to overestimation of treatment effects. The new formulae may be used to help researchers judge the possible extent of bias in past studies.

Keywords: survival data; risk heterogeneity; frailty; informative censoring; Proportional Hazards model; causal effect; Randomised Controlled Trials; Accelerated Failure Time model; G estimation; bias.

1. Introduction

This paper is concerned with estimation of treatment effect size from Randomised Controlled Trials (RCTs) using survival data, that is, when the outcome of interest is T_1 , time from randomisation to occurrence of a health-related event, E . In RCTs, patients are randomised to treatment groups; an attraction of randomisation is that it helps achieve balance between randomisation groups on known and *unknown* risk factors. Use of randomisation implies a concern for unmeasured heterogeneity in risk. The semi-parametric Proportional Hazards (smPH) method[1] is often used to estimate treatments effects in the form of Rate Ratios (RR). In its simple form, it assumes risk homogeneity across all individuals throughout follow-up; in the stratified version, the assumption applies within strata defined by covariates. The focus here is on what happens when there is *unmeasured heterogeneity* (UH), i.e. heterogeneity not captured by the covariates used in the analysis, under two forms of censoring.

It is often presumed that randomisation will prevent bias in estimation of treatment effects. But, a paper on omitted covariates in randomised designs concluded that “one cannot rely on randomization to guarantee consistent estimates of treatment effect for nonlinear response variables”[2]. A series of work[3-7], often referred to under the key word ‘frailty’, explains why this is so for survival data given UH. Those at highest risk, i.e. the most vulnerable to E , will succumb earlier in time than those less frail; thus, as events E accumulate, the average vulnerability of survivors decreases progressively. When there is a differential treatment effect, the magnitude of this selection effect will vary between treatment arms. It can be shown [7,8] that the apparent RR at time t will be attenuated compared to the true causal effect of treatment; attenuation increases with t . An expression for the bias in RR estimation when using a simple smPH model in this situation has been derived[9], but it and its approximation form generally require numerical evaluation of an integral.

Here we give a new approximation which is very simple to evaluate, and which emphasizes the importance of the hazard rate for bias size. Furthermore, it will also be shown that this bias is present and of the same magnitude using simple parametric Accelerated Failure Time (AFT) models[10,11] which characterise group differences by a survival time Acceleration Parameter (AP) and which some[8,11,12] have suggested may perform better than smPH. In contrast, there is no bias when using G estimation[13,14], a semi-parametric AFT method, under *simple censoring* (defined below).

Data from individuals where T_1 is not observed are called censored. An RCT protocol should define, in advance, rules for the maximum observation period $(0, M_i)$, say, for each enrolled individual i , $i=1, \dots, n$. In practice, M_i is often determined by the combination of a fixed study end date with a variable date of enrolment. Here, *simple censoring* is said to occur when observation continues to a *pre-determinable* M_i with no event observed. We say that there is *additional censoring* if observation ceases before this time, with no event observed. Additional censoring may or may not be *informative*. Censoring at time c is said to be *uninformative* if a patient whose survival time is censored at c is representative of all patients who have survived to c [15]; this is true of simple censoring. When covariates are used, the property needs hold only within covariate strata. An example of *informative* censoring is when drop-out from a study, with subsequent loss to follow-up, is linked somehow to the health status associated with risk of E . The validity of common methods for survival analysis depends on censoring being uninformative. Previous work on UH assumes simple censoring.

It is well known that estimation of RR from simple analyses given IC may be biased[16-18]. Censoring may be viewed as a competing risk; the fundamental problem is that, in general, one cannot recover the distribution of E , had C been absent, unless C is uninformative or one is willing to make extra assumptions[19]. Various sensitivity analyses (eg [16,18]) and estimation methods based on new assumptions (eg[17]) have been proposed to deal with it.

Here we approached the bias problem in a new way, extending the understanding of selection effects when there is UH under simple censoring. Note that if patients are homogeneous with respect to risk of E within analysis strata, then censoring *cannot* be informative. Thus, to say that there is informative censoring (IC) is to imply UH. A simple expression for the bias in *causal* (see later) Rate Ratio estimation under IC is derived. This shows two influences – one that would be present even without IC because of heterogeneity, the other to do with IC; these may act in different directions and so may partially cancel each other. Thus, we show that under IC the study bias can actually be *less* than if there was simple censoring only, and can even be zero. The focus on bias relative to the *causal* RR distinguishes this work from some others[16,17] where bias seems to be defined differently (see Discussion).

A *causal effect* is a contrast between potential outcomes under a treatment and control for a given patient[[20-21]. Such ‘individual level’ measures have direct clinical relevance since they state what would change if the patient was treated; they should therefore be the focus of analysis in a treatment trial. Individual causal effects are not usually estimable, but it is easy to show that, for a difference measure of effect, the difference between group means is an unbiased estimator for the corresponding patient average causal effect. Here, the focus is on estimation of the causal Rate Ratio, a ratio of potential risks, with the assumed aim to estimate the patient average of this measure. As implied earlier, some analytical approaches, eg smPH, may not estimate it correctly when there is UH.

A motivation for the present work stems from the fact that the vast majority of clinical trial reports are based on the smPH method[22], even though models which acknowledge heterogeneity (so called ‘frailty models’) exist. A bias formula for the smPH estimator under UH could help researchers judge whether published treatment effect estimates might be susceptible to appreciable bias, which task might inform a meta-analysis, for example. As will become clear, it might also be used to delineate situations in which bias from models

which assume homogeneity is acceptably small and therefore simple models would suffice. Another motivation is that many threats to valid causal effect estimation in trials with survival data – including IC – implicitly assume a correlation between behaviour and underlying risk and therefore, as argued above, bias is the net effect of two phenomena. Our approach to IC is potentially extendable to understand these problems too.

Whilst, as shown later, it is possible to ‘test’ for the presence of heterogeneity, this approach to analysis is not particularly advocated here. The paper takes the view, argued in Section 2, that heterogeneity should be the presumption. Thus, while it is true [23] that a model which assumes homogeneous risks but a declining Rate Ratio may fit the data as well as a model which allows heterogeneity but assumes a time invariant effect, here the a priori position is that the heterogeneous model is preferable. After Section 2, the paper proceeds as follows. In Section 3, previous results[7] on selection effects under *simple censoring* are used to derive an approximate formula for the expected bias in RR estimation using smPH. For the case of *informative censoring*, new formulae are derived for time dependent selection effects and the expected bias in smPH analysis. In Section 4, other analysis methods based on Accelerated Failure Time (AFT) models are reviewed from the heterogeneity perspective; these include AFT models which allow for UH in risk (‘frailty models’) and G estimation[13, 14], a semi-parametric method. In Section 6, the behaviours of smPH and AFT methods, including the frailty models, are evaluated against simulated data with either simple or informative censoring. The predictive abilities of the new formulae derived here are also assessed against these data. Simulation study methods are described in Section 5.

Section 7 compares the new formula for bias under simple censoring with a previous expression[9] for the bias in RR estimation. In Section 8, results from two published RCTs – one concerned with colorectal cancer[24] and the other with lung cancer[25] are re-evaluated assuming that there is unmeasured patient heterogeneity. The Discussion reviews

assumptions and applications. For simplicity, the main results assume no covariates in the analysis except for treatment group.

2. Homogeneity and heterogeneity: definition and reality

The time aspect of the homogeneity assumption implicit in the smPH method needs emphasis. Suppose firstly that there will be no covariates in the analysis. Let $M_{\max}$ be the maximum value of M_i where M_i is defined above. The patient population is said to be homogeneous in terms of risk if, in the absence of treatment effects, *at each time point during*

$0 < t \leq M_{\max}$, the hazard rate for E at time t, ie $\lambda^E(t) = \lim_{\delta t \rightarrow 0} \frac{P\{t \leq T_1 < t + \delta t \mid T_1 \geq t\}}{\delta t}$ *is the*

same for all survivors. If the analysis is to be stratified by additional covariates, a similar definition applies but within each stratum. This definition is implicit in smPH analysis because its log partial likelihood assumes that, just prior to each event, surviving patients are essentially interchangeable (within treatment or other strata), and this interchangeability continues right up until the last event.

This assumption seems highly implausible for most disease and health outcomes, with the possible exception of accidents. Even assuming that a completely homogeneous group of patients entered the trial at $t=0$, the homogeneity assumption contradicts the idea of disease occurrence (or recovery) as a *cumulative effect* of processes evolving over time: with this assumption, only ‘chance’ and the treatment decide who will be next. Consider the debate concerning the extent to which (re)occurrence of cancer is a question of “bad luck” arising from “random mutations” during DNA replication of stem cells[26-28]. The homogeneity assumption would require not only a belief that all mutations are down to bad luck but that there must be no cumulative effect of past mutations – which is implausible. Accumulation is consistent with the view expressed elsewhere[29], that some heterogeneity is always present.

The practical question is how much heterogeneity. A recent paper[29] concerned with evidence of heterogeneity in lifetime disease risk is noteworthy as to methods of investigation and results. Apparently, the size of the ‘familial Relative Risk’ - a measure of increased risk associated with having a family member who already has the disease - can be used to estimate between family risk heterogeneity. Evidence from studies focussing on cancer suggests 20 or 40-fold differences between the highest and lowest deciles of risk. When heterogeneity was expressed as a Gamma distributed risk multiplier, Z , with mean 1, values of $V(Z)$ of 1 or more were implied. Some studies suggested zero susceptibility to some cancers (eg colon[30]) in a proportion of the population. Elsewhere, for outcomes as diverse as tooth filling decay[31] and myocardial infarction[6], values for heterogeneity, expressed in terms of $V(\ln Z)$, of 0.9-1.1 were reported. Certainly one might expect that, in analysis which include covariates, the extent of heterogeneity within strata might be less than found in these reports; also risk heterogeneity among patients with a given disease may be less than in the general population. Nevertheless, this limited evidence suggests that risk heterogeneity be taken seriously in RCTs, especially primary prevention trials in the general population.

3. Bias in smPH effect estimation from heterogeneous populations

3.1 Simple censoring

Firstly assume there is a single population with no treatment intervention and simple censoring only. Let the hazard rate of E for patient i at time t be $\lambda_i^E(t) = z_i^E \lambda^E(t)$, $t > 0$, $i=1, \dots, n$, where z_i^E is an unobserved patient characteristic with mean 1 and variance θ^E .

Although it is convenient to describe variation in z_i^E using established statistical distributions for a random variable, Z^E , say, here the distribution is considered a fixed aspect of a given population. Homogeneity in risk implies $z_i^E = 1$ for all i ; the cumulative hazard and survivor

function under homogeneity are denoted by $H^E(t) = \int_0^t \lambda^E(u) du$ and $S^E(t) = \exp\{-H^E(t)\}$

respectively. For the heterogeneous case, the expected values of the hazard among survivors

at t and the other functions are denoted by $\lambda_\theta^E(t) = E[\lambda_i^E(t) | T \geq t]$, $H_\theta^E(t)$ and $S_\theta^E(t)$

respectively. Expressions for $\lambda_\theta^E(t)$ under various distributional assumptions for Z^E have

been derived[3-6,32]; the one-parameter Gamma and Inverse Normal distributions are often

cited. For these distributions with mean 1 and variance θ^E , it can be shown that:

$$\text{Gamma:} \quad \lambda_\theta^E(t | \text{simple censoring}) = \lambda^E(t)[1 + \theta^E H^E(t)]^{-1} \leq \lambda^E(t) \quad (1)$$

$$\text{Inverse Normal} \quad \lambda_\theta^E(t | \text{simple censoring}) = \lambda^E(t)[1 + 2\theta^E H^E(t)]^{-1/2} \leq \lambda^E(t) \quad (2)$$

In both, $\lambda_\theta^E(t) / \lambda^E(t)$ decreases with t . When $\theta^E H^E(t)$ is small, (1) and (2) are approximately

equal but the Gamma expression decreases more rapidly with t . The corresponding survivor

functions are

$$\text{Gamma:} \quad S_\theta^E(t) = [1 + \theta^E H^E(t)]^{-1/\theta} \quad (3)$$

$$\text{Inverse Normal:} \quad S_\theta^E(t) = \exp\left[\left(1 - \{1 + 2\theta^E H^E(t)\}^{1/2}\right) / \theta^E\right] \quad (4)$$

The focus of the new results here will be on the Gamma distribution but the approach easily

extends to other choices.

Now consider a RCT where patients are randomly allocated to a control ($x_i=0$) or a treated

group ($x_i=1$) at $t = 0$. Assume the analysis aims to estimate the causal Rate Ratio which is the

ratio of potential hazard rates under treatment and control for each patient and that the ratio is

time invariant and the same, $\exp(\beta^E)$, for all patients. Hence rates are $\exp(\beta^E x_i) \lambda_i^E(t)$,

$i=1, \dots, n$. In contrast, from (1) the ‘Survivor Rate Ratio’ at t , SRR say, is[7]:

$$\text{Gamma: } SRR(t | \text{simple censoring}) = \frac{\lambda_{\theta}^E(t | x=1)}{\lambda_{\theta}^E(t | x=0)} = \exp(\beta^E) \frac{1 + \theta^E H^E(t)}{1 + \exp(\beta^E) \theta^E H^E(t)} \quad (5)$$

There is no bias when $\exp(\beta^E) = 1$. Otherwise, as the following expression shows, the bias is *toward the null* and, for fixed t , increases with each of $\exp(\beta^E)$, θ^E and $H^E(t)$:

$$\text{Gamma: } \frac{SRR(t) - 1}{\exp(\beta^E) - 1} = \frac{1}{1 + \exp(\beta^E) \theta^E H^E(t)}$$

Aalen[7] proposed an extended version of the Gamma distribution where a proportion of the population has zero risk and shows that $SRR(t)$ may eventually cross the null; although this distribution may have applications in selected RCT applications, we focus here on the simpler Gamma distribution which is sufficient to show the main drivers of bias.

3.2 Informative censoring

Again consider a single population with no intervention but now with additional censoring. Although there may be several causes of this, for present purposes it is sufficient to consider only their joint effect, namely occurrence of a second event, C, which prevents further observation of E. Event C may be viewed as a realisation of a second process with corresponding time to event, T_2 . Define the hazard function for C as $\lambda_i^C(t) = z_i^C \lambda^C(t)$,

$i=1, \dots, n$, where $\lambda^C(t) = \lim_{\delta t \rightarrow 0} \frac{P\{t \leq T_2 < t + \delta t | T_2 \geq t\}}{\delta t}$ and Z^C is a random variable with

mean 1 and variance θ^C . If $z_i^C = 1$ for all i , everyone has the same chance, $\lambda^C(t)$, of C, with

corresponding cumulative hazard function and survivor functions $H^C(t)$ and $S^C(t)$

respectively. *Informative censoring* (IC) requires heterogeneity in E and C risks and a non-

zero correlation, ρ say, between Z^E and Z^C . Henceforth we assume that the additional

censoring by C is informative, ie $\rho \neq 0$.

Two expressions for $\lambda_\theta^E(t)$ are derived here by specifying a Gamma joint density distribution, $f(z^E, z^C)$. There are several bivariate Gamma distributions[33]; but derivation of exact results for is made easier when the Laplace transform for the joint density takes a simple form. One convenient choice is a bivariate Gamma distribution defined by Chatelain[34] with Laplace transform,

$$L(s_1, s_2) = \iint e^{-s_1 z^E} e^{-s_2 z^C} f(z^E, z^C) dz^E dz^C = [1 + p^E s_1 + p^C s_2 + p^E p^C (1 - \rho) s_1 s_2]^{-q} \quad (6)$$

and marginal univariate Gamma (q, p^E) and Gamma(q, p^C) distributions respectively for Z^E and Z^C ; q is a Gamma shape parameter and p^E and p^C are scale parameters. Because we need marginal means of 1, we have to set $p^E = p^C = 1/q$ which then implies equal variances for Z^E and Z^C of magnitude $1/q = \theta$, say. In the Appendix, the following result for the selection effect at t with this bivariate Gamma (BG 1) is derived:

$$\text{BG 1: } \lambda_\theta^E(t | \text{inf censoring}) = \lambda^E(t) \left[1 + \theta H^E(t) + \frac{\rho \theta H^C(t)}{1 + (1 - \rho) \theta H^C(t)} \right]^{-1} \quad (7)$$

The BG1 assumption of equal variances is restrictive. In the Appendix we also consider a different bivariate Gamma which has univariate Gamma means of 1, different variances with $\theta^E > \theta^C$, but now ρ is constrained to be $\sqrt{\theta^C / \theta^E}$. The corresponding result is:

$$\begin{aligned} \text{BG 2: } \lambda_\theta^E(t | \text{inf censoring}) &= \lambda^E(t) \left[1 + \theta^E H^E(t) + \theta^C H^C(t) \right]^{-1} \\ &= \lambda^E(t) \left[1 + \theta^E H^E(t) + \rho \sqrt{\theta^E \theta^C} H^C(t) \right]^{-1} \end{aligned} \quad (8)$$

The constraint on ρ seems even more restrictive but (8) serves to show the influence of θ^C , and suggests that θH^C in (7) might be replaced by $\sqrt{\theta^C \theta^E} H^C$ to give an approximate result for a similar distribution without the equal variance assumption or restrictive ρ . Other

bivariate distributions, including further Gamma choices[33], are possible but these two results are sufficient to suggest some general points: there are two components to the selection effect, one due to E events and one coming from C events; the direction of impact of the latter depends on the sign of ρ .

Returning to the 2-arm RCT context, we now allow treatment to potentially impact C; the impact on E is as before. Assume that treatment ($x=1$) multiplies the risk of IC compared to ‘control’ by $\exp(\beta^C)$; hazard rates for C are $\exp(\beta^C x_i) z_i^C \lambda^C(t)$, $i=1, \dots, n$. For example, increased likelihood of dropouts due to treatment side effects would mean $\exp(\beta^C) > 1$ while increases in the control group due to lack of perceived improvement perhaps, would mean $\exp(\beta^C) < 1$. It follows from (7) for BG1 that the ‘Survivor Rate Ratio’ for E at time t is now:

$$SRR(t | \text{inf censoring}) = \exp(\beta^E) \frac{\left[1 + \theta H^E(t) + \frac{\rho \theta H^C(t)}{1 + (1 - \rho) \theta H^C(t)} \right]}{\left[1 + \theta \exp(\beta^E) H^E(t) + \frac{\rho \theta \exp(\beta^C) H^C(t)}{1 + (1 - \rho) \theta \exp(\beta^C) H^C(t)} \right]} \quad (9)$$

Consider how this differs from simple censoring (5) *at a fixed t* , when $\rho > 0$.

- (i) Now there can be bias even when $\exp(\beta^E) = 1$: $SRR(t) > 1$ when $\exp(\beta^C) < 1$ and vice versa; only if also $\exp(\beta^C) = 1$, is $SRR(t) = 1$.
- (ii) When rates of informative censoring are the same for both groups, ie $\exp(\beta^C) = 1$, but $\exp(\beta^E) \neq 1$, $|SRR(t) - 1|$ is actually closer to $|\exp(\beta^E) - 1|$ than under simple censoring.
- (iii) In general, the impact of informative censoring on SRR depends on whether $\exp(\beta^C) - 1$ has the same sign as $\exp(\beta^E) - 1$. When they have the same sign and $\rho > 0$, the bias of $SRR(t)$ towards the null tends to be increased, although not always. But if they have different signs, informative censoring *reduces* the attenuation and might even eliminate it at some t .

These patterns of dependence on $\exp(\beta^C)$ are reversed when ρ is negative. Of note, now bias away from the null, ie $|SRR - 1| > |\exp(\beta^E) - 1|$, can occur if $H^C(t)$ is large enough compared to $H^E(t)$; when $\exp(\beta^E) < 1$, we also need $\rho(\exp(\beta^C) - 1) > 0$ but when $\exp(\beta^E) > 1$, we need $\rho(\exp(\beta^C) - 1) < 0$ for this to occur. Figure 1(a) illustrates the dependence of $SRR(t)$ on $\exp(\beta^C)$ given $\exp(\beta^E) = 0.5$ and $\theta = 1$, when either $\rho = -0.5$ or $\rho = 0.5$; $SRR(t)$ under simple censoring is also shown. At the chosen point in time, $H^E(t) = H^C(t) = 0.5$. Figure 1(b) shows the same scenarios but with $H^E(t) = 0.1$; situations where $SRR(t)$ is biased away from 1 are shown.

insert Figure 1 here

3.3 Bias in a single RR estimate under simple censoring and under IC

Equations (5) for simple censoring and (9) for IC are functions of time: they do not show the expected bias when a single RR is estimated for the whole study time span. We propose that, for simple censoring, the RR estimator from the smPH method, $\exp(\beta_{smPH}^E)$, say would be close to (5) with t set at t_{med} , *the median study event time for all groups combined*, i.e.

$$\text{Gamma: } \exp(\beta_{smPH}^E | \text{simple censoring}) \approx \exp(\beta^E) \frac{1 + \theta H^E(t_{med})}{1 + \exp(\beta^E) \theta H^E(t_{med})} \quad (10)$$

Likewise, under IC, a corresponding expression based on (9) but allowing different θ^E and θ^C (see above) is proposed;

$$\exp(\beta_{smPH}^E | \text{IC}) = \exp(\beta^E) \frac{\left[1 + \theta^E H^E(t_{med,IC}) + \frac{\rho \sqrt{\theta^E \theta^C} H^C(t_{med,IC})}{1 + (1 - \rho) \sqrt{\theta^E \theta^C} H^C(t_{med,IC})} \right]}{\left[1 + \theta^E \exp(\beta^E) H^E(t_{med,IC}) + \frac{\rho \exp(\beta^C) \sqrt{\theta^E \theta^C} H^C(t_{med,IC})}{1 + (1 - \rho) \exp(\beta^C) \sqrt{\theta^E \theta^C} H^C(t_{med,IC})} \right]} \quad (11)$$

where $t_{med,IC}$ = the median E event time given IC. The logic behind (10-11) is as follows. In a smPH analysis, the log partial likelihood is a sum of terms, one for each E event time; at

each event time, $SRR(t)$ changes as $H^E(t)$, and also $H^C(t)$ where applicable, increases. Since $\hat{\beta}_{smPH}^E$ is found by maximizing the sum of these likelihoods, we might expect the corresponding estimand to be close to $SRR(t)$ at the middle event time.

Note that $H^E(t_{med\ IC})$ in (11) will be less than $H^E(t_{med})$ in (10) because the number of E events is reduced by IC; hence a simple comparison of (10) and (11) assuming the same t_{med} would be misleading. This reduction in the E event rate, on its own, would tend to decrease bias. But IC also changes the selection imbalance between groups in terms of the survivor $\lambda_i^E(t)$ distributions, possibly favourably. The joint effect of these factors is illustrated later.

4. Alternatives to the smPH model for treatment effect estimation

Several authors[8,11,12] have suggested that an Accelerated Failure Time (AFT) model might perform better under heterogeneity than the Cox PH. The homogeneous AFT model for two groups assumes that $S(t | x) = S(\exp(\beta^{*E}x)t)$ where β^{*E} is the Acceleration Parameter (AP); $\exp(\beta^{*E})$ measures deceleration/acceleration of survival time for the $x=1$ group versus $x=0$. In AFT models implemented in most statistical software, the distribution across time for $\lambda^E(t)$ must be specified with the Weibull a common choice. This has a useful property in that events satisfy both a PH model, $\lambda^E(t | x) = \lambda_0^E(t) \exp(\beta^E x)$, and an AFT model, with $\beta^{*E} = -\beta^E/\kappa$ where λ and κ are the Weibull scale and shape parameters respectively. The Weibull AFT can also be expressed as a regression model of the form $\ln T = \alpha + \beta^{*E}x + \kappa^{-1}\varepsilon$ where ε has an extreme value distribution. When fitted via maximum likelihood (ML), both β^{*E} and κ are estimated, thus providing another potential route to estimating β^E .

One reason for the favourable view of AFT models is that if $S(t | x) = S(\exp(\beta^{*E}x)t)$, then also $S_\theta(t | x) = S_\theta(\exp(\beta^{*E}x)t)$, ie the AFT assumption still holds under heterogeneity[12].

However it does not follow that a parametric, homogeneous AFT model will be trouble free, because the time distribution assumption will be wrong, eg we might specify a Weibull survivor function for $S(t)$, whereas $S_{\theta}^E(t)$ is (3), which is not Weibull. Therefore bias in β^{*E} and/or κ estimation should still be expected.

AFT models can be fitted without a time distribution assumption using G estimation[13,35]. In this method, a *potential* event time, U_i , under control treatment is defined for everyone in the trial, regardless of treatment arm; U_i is related to the observed times t_i and unknown β^{*E} by $U_i(\beta^*) = t_i \exp(-\beta^{*E} x)$. The method also uses the predefined times M_i to define a new censoring rule on the U scale. Under randomisation, there should be no systematic difference between treatment arms in the distribution of U; therefore β^{*E} is estimated by the value $\hat{\beta}^{*E}$ which gives a zero test (eg log rank) statistic when comparing $U_i(\hat{\beta}^{*E})$ between randomisation arms. We expect this method to be bias free for β^{*E} under simple censoring but not necessarily under IC.

The STATA software package allows parametric AFT ‘frailty’ models to be fitted with risk heterogeneity specified by either Gamma or Inverse Normal ‘frailty’ distributions; the likelihood uses the survivor functions (3-4), and θ^E is estimated. We might expect these methods to perform well under simple censoring but it is not obvious what happens under IC.

5. Simulation study methods

The aims of the simulation study were, for simple censoring and IC scenarios, to (i) check the accuracy of (10-11) for prediction of bias from homogeneous models and (ii) evaluate the performance of various smPH and AFT estimation methods, including frailty models, for β^E or β^{*E} . The simulation studies mimicked a two-armed randomised trial among a fixed

population of 1000 patients with equal probability of being randomly assigned to either a treatment arm or control arm. The trial duration was 24 months with $M_i = 24$ for all.

Parameter values were chosen purposefully, using (10) as a guide, to give discernible bias in the smPH method, and to give a high degree of IC when applied. These goals require high event rates and high values of θ and ρ . Thus, following Michael et al[36], values of z_i^E and z_i^D for a population of size 1000 were generated once from a bivariate Gamma distribution with the aim of achieving means of 1, variances of 1 and correlation of 0.5; the actual distribution used had means of 1, $\theta^E = 0.971$, $\theta^C = 1.000$ and $\rho = 0.503$. This aspect of the population was fixed across all simulations. Event rates, $\lambda^E(t)$ and $\lambda^C(t)$ in the untreated arm, were assumed to follow Weibull distributions for which the shape parameter, κ , took values 0.75, 1 or 1.5 in different simulation sets. The corresponding Weibull scale parameters were set at $0.6/24^\kappa$ per month, so that $H^E(24) = H^C(24) = 0.60$ for all κ , implying that total expected numbers of events in the control group would be the same regardless of shape. Individual rates of E and C if untreated were then set at $z_i^E \lambda^E(t)$ and $z_i^C \lambda^C(t)$, $i=1, \dots, 1000$. The causal Rate Ratio $\exp(\beta^E)$ was variously set at 0.5, 0.75, 1, 1.5 and 2. For each $\exp(\beta^E)$ value, Rate Ratios for C were set at 0.5, 1, 2 or else simple censoring, $\lambda^C(t) = 0$, was assumed. Thus there were 20 simulations sets for each Weibull shape. For a given $\exp(\beta^E)$, values of the Acceleration Parameter β^{*E} vary with κ ($\beta^{*E} = -\beta^E/\kappa$); values of $\exp(\beta^{*E})$ are shown in Tables.

There were 1000 simulations in each set; the random number ‘seed’ for each set was identical across sets. Within a simulation, patients were allocated to active treatment with probability 0.5 using a draw from a Uniform (0,1) distribution. If allocated to treatment, the patient’s control rates for E and C were multiplied by $\exp(\beta^E)$ and $\exp(\beta^C)$ respectively. To generate the Weibull event times, Uniform random numbers were generated and transformed appropriately. Under simple censoring, there were no C events; E event times were censored

at $t = 24$. In the other scenarios, the minimum of $(24, t_{2i})$ where t_{2i} is the time of event C, was used to censor observation of E.

The smPH and AFT analyses were applied to each simulated data set using STATA software routines **stcox** and **streg**. G estimation was implemented with a user written, STATA software routine, **strbee**[35] assuming $M_i = 24$. Parametric AFTs were fitted assuming E times had a Weibull distribution, both with (HET) and without (HOM) allowance for heterogeneity. The AFT HOM parametric model was used to give both $\hat{\beta}^{*E}$ and $\hat{\beta}^E$, so that the latter could be compared with estimates from the smPH model. Estimates of β^{*E} from the AFT HET parametric model were contrasted with those from the G method. To summarise estimates of β^E , the set mean $\bar{\hat{\beta}}^E$ was calculated, then $\exp(\bar{\hat{\beta}}^E)$ since $\exp(\beta^E)$ is more meaningful in practice than β^E ; when $\beta^E \neq 0$, $(\bar{\hat{\beta}}^E - \beta^E) * 100 / \beta^E$ - referred to as Standardised Bias Percent (SB%) - was also found. Similar measures were used to summarise estimates of $-\hat{\beta}^{*E}$; the minus sign allows an equivalent meaning to that for β^E when $\kappa = 1$, while not affecting interpretation of the sign of SB% (see below). Mean values of $\hat{\kappa}$ across simulations are reported. The median E event time for the two groups combined in each simulation was found. The set mean of these median times was used to find $H(t_{med})$ which was then inserted into the bias formulae (10-11).

The power when using a 2-sided t-test, with 1000 simulations per set, is always 88% to detect a bias in estimation of β^E or β^{*E} of a size equivalent to $0.1 * SD(\hat{\beta}^E)$. As reported later, $SD(\hat{\beta}_{smPH}^E)$ was between 0.10 and 0.15 depending on β^E and β^C , but some other SDs reached 0.24. Methods for which a one-sided t test on $\hat{\beta}^E$ or $\hat{\beta}^{*E}$ gave a p value < 0.05 are judged to be biased. Results with $p < .05$ for which there is a negative SB% imply bias towards the null, while those with a positive SB% are biased away from the null, ie they exaggerate effect size.

6. Evaluation of formulae and methods

We firstly clarify expectations as regards the effect of κ on the performance of the smPH method and on results from the bias formulae (10-11). The influence of κ on timing of events in the simulations is shown in Table 1(a) via median event time, t_{med} or $t_{\text{med,IC}}$, and is as expected. However, the magnitude of event times does not enter into likelihood of the smPH method, suggesting that it might be invariant with respect to κ , all else being equal. In the simulations, $H^E(24)$ was fixed for all κ ; Table 1(b) shows that $H^E(t_{\text{med}})$ varied only trivially with κ . Thus, with $\exp(\beta^E)$, $\exp(\beta^C)$, θ^E , θ^C and ρ fixed, predictions from the formulae (10-11) are not expected to vary much with κ .

In fact, with other parameters held constant, $\hat{\beta}_{\text{smPH}}^E$, $i=1, \dots, 1000$ were identical across κ since, beside the above property of the smPH method, the ‘seed’ for starting the random number generation in the simulation sets had been identical across κ . Thus, results for bias in this method do not need to be shown separately for each κ . Bias is summarised in Table 2, with the set Standard Deviations of $\hat{\beta}_{\text{smPH}}^E$ given in Appendix Table A1. When $\beta^E \neq 0$, the smPH method is biased in all situations shown, both SC and IC. When $\beta^E = 0$, it is unbiased under SC and under IC only if $\exp(\beta^C) = 1$. As expected, bias is always towards the null under SC.

Under IC, the bias in β_{smPH}^E is sometimes *less* than under SC: for example, when $\exp(\beta^E)=1.50$ and $\exp(\beta^C)=0.5$, SB% is -2.4% compared to -14.7% under SC. This is in part because IC reduces the cumulative hazard but also because, as noted earlier, when β^C has a different sign from β^E and $\rho > 0$, IC can favourably improve the selection imbalance between groups among survivors. Under IC, there were also scenarios with bias away from the null: for example, when $\exp(\beta^E)=0.75$ and $\exp(\beta^C)=2.00$, $\exp(\bar{\beta}_{\text{smPH}}^E)$ was 0.74 compared to 0.79 with simple censoring. The way in which bias changes with $\exp(\beta^C)$ in these results suggested that, in

principle, there may be no bias under IC in some circumstances. To test this important proposal, an additional set of simulations (not in Table) was run with $\exp(\beta^E)=0.50$ and $\exp(\beta^C)=2.17$; $\exp(\hat{\beta}_{smPH}^E)$ was 0.503 with a non-significant t test: $p = 0.18$.

To find predictions for $\exp(\beta^E)$ and β^E , the fixed values of θ^E , θ^C and ρ , together with $\exp(\beta^E)$, $\exp(\beta^C)$ and $H(t_{med})$ from Table 1, were inserted into (10) or (11). The predicted SB% for β^E was then calculated. The predicted values of $\exp(\beta^E)$ and SB% of the prediction were the same across κ values to 3 decimal places (1 place for SB%); thus, these are shown only once in Table 2 for each $\exp(\beta^E)$ and $\exp(\beta^C)$ combination. Both prediction formulae performed very well; the absolute difference between actual and predicted SB% tended to be $< 1\%$.

Surprisingly, estimates of β^E from the AFT HOM method, as given directly by the STATA routine **streg**, also did not vary with κ and so, again, those results in Table 2 do not need to distinguish between κ values. The bias in this method is only slightly smaller than that of the smPH model: thus the bias formulae could also apply to β^E estimation for this method too.

The AFT HOM method was also used to estimate β^{*E} and κ . These results are summarised in Table 3 using $\exp(-\hat{\beta}^{*E})$, but now separately for each κ since $\beta^{*E} = -\beta^E / \kappa$; values of $\exp(-\beta^{*E})$ are shown for comparison. While β^{*E} varied with κ , SB% for β^{*E} did not, and t-tests also returned the same p value. Judged by SB%, estimation of β^{*E} is less biased than estimation of β^E (Table 2), at least under SC. There appeared to be no bias for β^{*E} for some combinations of $\exp(\beta^E)$ and $\exp(\beta^C)$; examples are $\exp(\beta^E) = 0.5$, $\exp(\beta^C) = 1$ and $\exp(\beta^E) = 2.0$, $\exp(\beta^C) = 0.5$. Estimation of κ is biased downwards throughout.

Table 4 summarises estimates of β^{*E} from the semi-parametric G estimation method and from the AFT HET model; $SD(\hat{\beta}^{*E})$ values are given in Table A2. Results from both

methods were very similar, although the AFT method tended to give slightly less bias and, in Table A2, its $SD(\hat{\beta}^{*E})$ tended to be slightly smaller. For both methods, under SC, no t- test returned $p < 0.05$ and therefore both are judged unbiased within the limits of detection. But, under IC, both are usually biased. One exception is the special case when $\exp(\beta^E) = 1$ and $\exp(\beta^C) = 1$; the smPH and AFT HOM were also unbiased for β^E in this scenario. Unforeseen when setting up the simulations for IC, the case where $\exp(\beta^E) = \exp(\beta^C)$, $H^E(t) = H^C(t)$ and $\theta^E = \theta^C$ represents a set of special conditions for performance of the HET AFT and G estimation methods: there is no bias. It seems that, in this case, the combined effect of the selection processes is equivalent to that from simple censoring alone but with variance θ replaced by $\theta(1+\rho)$; hence the model allowing risk heterogeneity is expected to remove bias. These special conditions are unlikely. Of more importance is the fact that both methods gave some large biases away from the null, for which SB% is positive; for example when $\kappa = 1$, $\exp(-\beta^{*E}) = 0.50$, $\exp(\beta^C) = 2.0$, $\exp(-\bar{\beta}^{*E})$ was 0.442 for the two methods.

For reasons given in Section 2, we do not particularly advocate testing of $H_0: \theta^E = 0$. However, in the interest of completeness, the performance of a 1 sided test[37] of $H_0: \theta = 0$, provided in STATA **streg** when AFT HET models are fitted, is reported. Recall $\theta^E = 0.971$. The test with level 0.05 rejected H_0 only between 30% and 70% of the time (more often for higher β^E when total trial events were higher). The low power for $\theta \neq 0$ is not a result of low power generally: in terms of events, simulated studies were of sufficient size to give power $\geq 89\%$ for a test to detect a 30% reduction in events using a 2-tailed test with $\alpha = 0.05$. Thus it seems that a strategy of testing for heterogeneity is likely to be underpowered in typical studies. Also, up to 12% of θ values in some simulation sets were effectively zero (i.e. < 0.0001) and there was a strong correlation (~ 0.82 - 0.85) between $\hat{\kappa}$ and $\hat{\theta}$ in simulation sets. It seems as if the separation of their influences is difficult, perhaps because both parameters

contribute to the variance. However the results here for testing $H_0: \theta = 0$ are indicative only; we do not claim this to be a full assessment of the power of this or similar tests {Claeskens, 2008 279 /id}, which might require a different simulation design with variation in degree of heterogeneity.

7. Comparison with Henderson and Oman approximation

Henderson and Oman[9] derived a general expression for β_{smPH}^E from a simple PH model under UH, then went on to seek simpler expressions under three forms of censoring (not including IC). Only in the case of no censoring, with a Gamma distributed Z , was a closed form expression found: this stated $\beta_{smPH}^E / \beta^E \approx (1 + \theta^E)^{-1}$. However it is highly unlikely that patients in an RCT would be followed long enough to observe events in all; thus, this expression is mainly of theoretical interest. The other scenarios were a form of ‘random censoring’ – linked to the unspecified baseline survivor function but not to Z - and censoring with a predetermined time, M . In neither case was the expression easy to interpret, thus they evaluated numerically both exact β_{smPH}^E / β^E and its approximation in 27 settings for the random censoring scenario. These comprised three expected censoring fractions (0.2,0.5,0.8), three $SD[\ln(Z)]$ values - corresponding to θ^E values of 0.2232, 1.2325 and 2.2933 of a Gamma distributed Z , and β equal to 1, 2 or 4. The approximation was very good when $\beta^E = 1$ or 2, less so when $\beta^E = 4$, but this last value ($\exp(\beta^E)=55$) seems less important. Their results for $\theta^E = 1.2325$ are reproduced in Table 5(a) in the original form; for fixed θ^E and β^E , bias increases as the censoring fraction decreases. Results for censoring with a predetermined M were said to be ‘qualitatively similar’; the formula for the approximation in this case is:

$$\beta_{smPH} / \beta \approx (1 - E[M^Z])^{-1} \left\{ \int_0^{-\ln(M)} u \frac{E[Z \exp(-uZ)]^2}{E[\exp(-uZ)]} du - E[ZM^Z] \ln(M) \right\}. \quad (12)$$

Comparison between Table 5(a) and Table 2 is not possible given lack of setting overlap. In our tables, censoring fractions varied with β^E . To provide some comparison and to evaluate sensitivity to different forms of non-informative censoring, new simulations were generated with $\theta^E = 1.2325$, $\beta^E = -0.5, 0.5, 1, 2$ and giving average censoring fractions of 0.2, 0.5 or 0.8 (achieved by trial and error testing various $\lambda^E(t)$ values). These has $M = 24$ and Weibull hazards with $\kappa = 1.50$. To give increased precision for β_{smPH}^E / β^E estimation, there were 5000 simulations per set. Estimated β_{smPH}/β and predictions from (10) are shown in Table 5(b).

Within each setting (a) and (b), it is interesting that, if the censoring fraction is fixed, β_{smPH}^E / β^E does not vary much with β^E , at least within the range illustrated. But β_{smPH}^E / β^E from the two settings did not agree in examples with censoring fractions of 0.2 or 0.8. We conclude that their[9] formulae are sensitive to the type of censoring; thus their illustrated values of β_{smPH}^E / β^E cannot be taken as universal for a given β^E and censoring fraction. The approximation (10) is not tied in to a particular form of censoring. Taking these results and those in Table 2 together, when $-0.693 \leq \beta^E \leq 1$, its performance is very good (SB% no more than 1%), but it seems to underestimate bias when $\beta^E = 2$ ($\exp(\beta^E) = 7.4$).

8. Using formula to assess potential for bias in previous studies with simple censoring

In the two trials reported here there appears to have been no loss to follow up, which is often true when outcomes can be ascertained independently through a registry. RR estimates were reported based on simple analyses assuming homogeneity and with no covariate adjustment. As shown in (10), the determinants of bias are $H^E(t_{med})$ in the control group – which can be estimated - and θ , the rate variance. There are no direct estimates of θ from either study but external information might be used to inform ‘guesstimates’ or a range of values considered.

The first study evaluated a sigmoidoscopy screening programme for prevention of colorectal cancer in 55,736 Norwegians[24] with incidence at 5 years stated as the primary endpoint, although incidence based on 6-8 years follow-up was reported; by then there was a total of 485 events. From data in the published paper, t_{med} was around 3.3 years with $H(t_{\text{med}})$ in the control group estimated as 0.0044. For this general population there could be zero colorectal cancer risk in a proportion[30] suggesting high variability. If we take $\theta = 1$, as suggested by some studies (see Section 2), and a true RR of 0.70 - the effect size used for sample size calculations - equation (10) suggests that, in expectation, $\exp(\hat{\beta}_{smPH}^E)$ would be 0.701. This is a trivial bias. Even if θ was as large as 5, the expected value is only 0.704 and the conclusion unchanged.

The second study is a two arm trial of thoracic radiotherapy in lung cancer patients, with primary endpoint survival at one year, by which time there had been 367 events in total among the 495 patients. From Figure 2 of the associated paper[25], the median event time was estimated as 6 months and control group $H(t_{\text{med}})$ was roughly 0.35, close to pre-study estimates. The authors also reported an analysis based on 2 year survival with 478 events in total; for this analysis $H(t_{\text{med}})$ in the control group was roughly 0.48. Variability in this patient group might be expected to be lower than in the general population but we have no direct estimates; a θ value of 0.5 does not seem excessive but values of 1 and 5 are also considered. With $\theta = 0.5$, the critical product $\theta H(t_{\text{med}})$ is 0.24 for the longer follow-up, which is close to that in the simulations here with $\exp(\beta^E)=0.75$. Thus, from Table 1, we see that with $\exp(\beta^E)=0.75$, we expect $\exp(\hat{\beta}_{smPH}^E) = 0.788$, which seems an important bias. With $\theta=1$ ($\theta = 5$), this becomes 0.816 (0.911). The contrasting implications for the two studies, if θ was the same in each, show the importance of $H(t_{\text{med}})$.

9. Discussion

The new approximation (10) for simple censoring has been shown to be accurate for prediction of bias, over a range of β which is typical of RCTs, when using a smPH model or AFT HOM Weibull model to estimate a causal Rate Ratio when the underlying survival distribution is Weibull. Since the derivation of (10) made no assumption about the form of the survival distribution, there is no reason to expect that it would not perform equally well to predict smPH performance with other distributions. (On the other hand, for bias in parametric AFT HOM models with an assumed distribution different from the actual, there might be more divergence.) It is based on a Gamma distribution but the corresponding formula for the inverse Normal distribution is easily found from (2), and the general approach can be adapted to others. The formula is very simple to use compared to (12) say and is unlikely to be dependent on the type of censoring provided it is uninformative.

Of course, the need for a bias formula would be circumvented if appropriate analyses which allow for heterogeneity were to become the norm but, as noted in a 2010 publication[22], the vast majority of trial reports are based on the smPH method. Thus the formula may be of value for a qualitative reassessment of previous work of this type, as illustrated in Section 8, but also to those contemplating a new analysis. For example, in a general population screening trial, $H(t_{\text{med}})$ is likely to be very small: thus one might conclude that a simple homogeneous smPH analysis will be adequate.

This work assumes a constant causal treatment effect, β , and attributes any apparent change over time in $SRR(t)$ to selection. In reality, β might also decrease in addition to the presence of risk heterogeneity or, some may believe, there might be a decreasing β and no heterogeneity. In all these cases, a homogeneous hazards model or AFT model which allows β to be a function of t may fit the data well. Because of this, some authors[23] take the view

that there is no need for ‘frailty’ models. But this opinion presumes that model fit is the only determinant of choice of model. It is suggested here that a priori knowledge and beliefs – eg about the plausibility of the homogeneous assumption or a declining causal treatment effect – does and should influence model choices; data is rarely the sole criterion.

Formula (11) for IC is also shown to be accurate in predicting bias. Previously, Siannis et al[16] gave quite general approximate formulae for bias in rates, which in principle cover the correlated random effects scenario although allowing only equal rates of IC in both groups, ie $\beta^C = 0$. These were said to be accurate only when the magnitude of correlation between T and C ≤ 0.3 . But it seems that, there, bias was defined in relation to the (biased) estimates when there is heterogeneity but no IC: the formulae assume that the estimate with a zero correlation is unbiased which, as we have seen, is not true for a causal RR. Huang[16] proposed an analysis for overcoming IC where heterogeneity in $\lambda^E(t)$ and $\lambda^C(t)$ was due to a common random effect with weight 1 for E and α for C; the sign of α , has the same role as that of ρ here. Among other conclusions, it was stated that, if IC is ignored, the treatment log RR “ $\beta^{(T)}$ ” is overestimated when $\alpha > 0$ and $\beta^{(C)} > 0$. But, as shown here, this is not true in general of the causal parameter β^E . Again it seems that bias is defined in relation to the (biased) estimate of β^E when there is no IC.

Both (10) and (11) could also be useful for methodologists concerned with other threats to valid causal effect estimation, eg treatment non-compliance, in trials with survival data. It is known[39-42] that, regardless of outcome type, simple methods of adjustment for non-compliance will be biased when non-compliance is correlated with pre-treatment traits. For survival data, by analogy with the approach to IC, we can say the overall bias for $\exp(\beta^E)$ in this scenario will be a net effect of UH and non-compliance, and (10) is then useful for estimating the effect of UH alone.

Formula (11) is general in some aspects: ρ is unrestricted and there are no assumptions about the distribution of $\lambda^C(t)$. IC was modelled by assuming that untreated patients have different baseline propensities, z_i^C , to drop out, and Z^C is correlated with Z^E . This approach has been used by others as an example of how a joint distribution for T and C might come about (eg in [16]) and is not unrealistic: Z^C could reflect differences in several personality traits (eg persistence, helpfulness, sociability). The assumption that the propensity to drop out might be altered by membership of the treated group, ie $\beta^C \neq 0$, is plausible although not perhaps that it is uniformly altered. The Gamma distribution assumption for Z^C is restrictive although Gamma encompasses a number of different shapes. Nevertheless, we suggest that the main structure of the formula – showing the different components and how they combine – might not be sensitive to the choice of distribution.

This structure adds to our understanding of the impact of IC in concrete ways. Firstly we reiterate that IC acts in a similar way to risk heterogeneity by changing the distribution of $\lambda_i^E(t)$ in survivors. It also reduces $H^E(t_{med})$ which, alone, acts to reduce attenuation bias from heterogeneity. Thus, at first sight paradoxical, less bias in estimation of β^E than under simple censoring or even zero bias, is explainable. This paper also adds to knowledge by showing how the best methods for dealing with heterogeneity bias under simple censoring (G estimation and AFT HET) can produce a bias quite different from the one expected using homogeneous models. These methods cannot therefore be recommended when there is IC.

The direction of change induced by IC alone in the distribution of $\lambda_i^E(t)$ is knowable since it depends on the sign of ρ and whether β^C has the same sign as β^E ; the sign of ρ should be inferable, either from questions to patients about reasons for dropping out or by comparing baseline health information for leavers and non-leavers. But the precise amount of change is

not predictable since θ^C would be difficult to estimate. The important role of $H^C(t_{med\ IC})$ should however be noted: for IC to have much impact the product $\rho H^C(t_{med\ IC})$ must be large.

Finally, use of the phrase ‘frailty models’ to denote statistical models which allow risk heterogeneity[7,22,29] should be contrasted with medical use. There, “frailty is a long established clinical expression” that denotes a “state of increased vulnerability to poor resolution of homoeostasis after a stressor event”[43] and a frailty model is a clinical prediction tool to identify such people. It is suggested that in statistical modelling the term “risk heterogeneity” is more informative as to purpose, less confusing and less open to dispute[44].

Funding

Although there was no funding for this work, the author wishes to acknowledge support from MRC grant G0601643 for other work which prompted the issues investigated here.

Data Accessibility There are no original data in this manuscript

Appendix: Derivation of mean Hazards for some bivariate Gamma distributions

Let $T = \min(T_1, T_2)$. For individual i given $T \geq t$, the probability of an event of any type at time t is $\lambda_i^E(t) + \lambda_i^C(t) = \lambda_i(t)$ say, and the corresponding survivor function is

$$S_i(t) = P(T \geq t | i) = \exp\left[-\int_0^t \{\lambda_i^E(u) + \lambda_i^C(u)\} du\right] = \exp[-z_i^E H^E(t) - z_i^C H^C(t)]$$

It follows that

$$S_\theta(t) = P(T \geq t) = \int \int \exp[-z^E H^E(t) - z^C H^C(t)] f(z^C, z^E) dz^E dz^C = L[H^E(t), H^C(t)]$$

where $L[s_1, s_2]$ is the Laplace transform of f .

$$\text{Also } f(z^E, z^C | T \geq t) = \frac{P(T \geq t | z^E, z^C) f(z^E, z^C)}{S(t)} = \frac{\exp\{-z^E H^E(t) - z^C H^C(t)\} f(z^E, z^C)}{L[H^E(t), H^C(t)]}$$

$$\text{It follows that } \lambda_\theta^E(t) = E[\lambda_i^E(t) | T \geq t] = \frac{-\lambda^E(t) L'[H^E(t), H^C(t)]}{L[H^E(t), H^C(t)]} \quad (\text{A1})$$

$$\text{where } L'(s_1, s_2) = \delta L(s_1, s_2) / \delta s_1$$

BG1: The Laplace transform for the bivariate Gamma distribution defined by Chatelain[34] is given in (6). Substituting this and its derivative into (A1) and setting $p^E = p^C = 1/q = \theta$, gives

$$\lambda_\theta^E(t | \text{informative censoring}) = \frac{\lambda^E(t) [1 + \theta(1 - \rho)H^C(t)]}{1 + \theta H^E(t) + \theta H^C(t) + \theta^2(1 - \rho)H^E(t)H^C(t)} \quad (\text{A2})$$

The denominator in (A2) can be written as $[1 + (1 - \rho)\theta H^C(t)][1 + \theta H^E(t)] + \rho\theta H^C(t)$; hence (7) follows.

BG2: Suppose that Z^E has a univariate Gamma (q_1, θ_1) distribution, i.e. with shape parameter q_1 and scale parameter θ_1 ; $E(Z^E) = \theta_1 q_1$; $V(Z^E) = \theta_1^2 q_1$. Then suppose that

$$Z^C = \theta_2 Z^E / \theta_1 + \omega$$

where ω is an independent Gamma variate with distribution $G(q_2, \theta_2)$ with $\theta_2 < \theta_1$. From the more general result by Furman[45] for multivariate Gamma distributions, it follows that Z^C is univariate Gamma with $E(Z^C) = \theta_2(q_1 + q_2)$; $V(Z^C) = \theta_2^2(q_1 + q_2)$ and the joint distribution of (Z^E, Z^C) is bivariate Gamma with Laplace transform

$$L(s_1, s_2) = \iint e^{-s_1 z^E} e^{-s_2 z^C} f(\alpha_E, \alpha_C) dz^E dz^C = [1 + \theta_1 s_1 + \theta_2 s_2]^{-q_1} [1 + \theta_2 s_2]^{-q_2} \quad (\text{A3})$$

and their correlation is $\sqrt{\theta_2 / \theta_1}$. Because we need $E(Z^E) = E(Z^C) = 1$, we set $q_1 = 1/\theta_1$ and $q_2 = 1/\theta_2 - 1/\theta_1$. Then $V(Z^E) = \theta_1$ say and $V(Z^C) = \theta_2$. We then substitute (A3) and its derivative in (A1) to derive $\lambda_\theta^E(t)$, giving:

$$\lambda_\theta^E(t | \text{informative censoring}) = \lambda^E(t) [1 + \theta^E H^E(t) + \theta^C H^C(t)]^{-1} \quad (\text{A4}).$$

REFERENCES

1. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society B* 1972; **34**:187-220.
2. Gail MH, Wieand S, Piantadosi S. Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. *Biometrika* 1984; **71**:431-444.
3. Vaupel JW, Manton KG, Stallard E. Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography* 1979; **16**:439-454.
4. Hougaard P. Frailty models for survival data. *Lifetime Data Anal* 1995; **1**:255-273.
5. Hougaard P. Life Table Methods for Heterogeneous Populations - Distributions Describing the Heterogeneity. *Biometrika* 1984; **71**:75-83.
6. Hougaard P. Survival Models for Heterogeneous Populations Derived from Stable-Distributions. *Biometrika* 1986; **73**:387-396.
7. Aalen OO. Effects of frailty in survival analysis. *Stat Methods Med Res* 1994; **3**:227-243.
8. Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010; **21**:13-15.
9. Henderson R, Oman P. Effect of frailty on marginal regression estimates in survival analysis. *J R Statist Soc B* 1999; **61**:367-379.
10. Cox DR, O'D. *Analysis of Survival Data*. 1st edition. Boca Raton London New York Washington: Chapman and Hall/CRC; 1984.
11. Keiding N, Andersen PK, Klein JP. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 1997; **16**:215-224.
12. Gutierrez RG. Parametric frailty and shared frailty survival models. *STATA Journal* 2002; **2**:22-44.
13. Robins J, Tsiatis AA. Semiparametric Estimation of An Accelerated Failure Time Model with Time-Dependent Covariates. *Biometrika* 1992; **79**:311-319.
14. Joffe MM, Yang WP, Feldman H. G-estimation and artificial censoring: problems, challenges, and applications. *Biometrics* 2012; **68**:275-286.
15. Collett D. *Modelling Survival Data in Medical Research*. 1st edition. Boca Raton London New York Washington: Chapman and Hall/CRC; 1994.
16. Siannis F, Copas J, Lu GB. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics* 2005; **6**:77-91.
17. Huang X, Wolfe RA. A frailty model for informative censoring. *Biometrics* 2002; **58**:510-520.

18. Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Stat Med* 2014; **33**:4681-4694.
19. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci U S A* 1975, **72**:20-22.
20. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *American Statistical Association* 2005; **100**:322-331.
21. Rubin DB. Estimating causal effects of treatments in randomised and non-randomised studies. *Journal of Educational Psychology* 1974; **66**:688-701.
22. Wienke A. *Frailty Models in Survival Analysis*. Boca Raton London New York Washington: Chapman and Hall/CRC; 2010.
23. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Statistics in Medicine* 2002; **21**:3219-3233.
24. Hoff G, Grotmol T, Skovlund E, Bretthauer M. Risk of colorectal cancer seven years after flexible sigmoidoscopy screening: randomised controlled trial. *BMJ* 2009; **338**:b1846.
25. Slotman BJ, van TH, Praag JO, Knegjens JL, El Sharouni SY, Hatton M, Keijser A, Faivre-Finn C, Senan S. Use of thoracic radiotherapy for extensive stage small-cell lung cancer: a phase 3 randomised controlled trial. *Lancet* 2015; **385**:36-42.
26. Couzin-Frankel J. The bad luck of cancer. *Science* 2015; **347**:12.
27. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 2015; **347**:78-81.
28. Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature* 2016; **529**:43-47.
29. Aalen OO, Valberg M, Grotmol T, Tretli S. Understanding variation in disease risk: the elusive concept of frailty. *Int J Epidemiol* 2015; **44**:1408-1421.
30. Svensson MH, Svensson E, Lasso A, Hellstrom M. Patient acceptance of CT colonography and conventional colonoscopy: prospective comparative study in patients with or suspected of having colorectal disease. *Radiology* 2002; **222**:337-345.
31. Aalen OO, Bjertness E, Soonju T. Analysis of dependent survival data applied to lifetimes of amalgam fillings. *Stat Med* 1995; **14**:1819-1829.
32. Vaupel JW: Inherited Frailty and Longevity. *Demography* 1988; **25**:277-287.
33. Nadarajah S GA. Some bivariate gamma distributions. *Applied Mathematics Letters* 2006; **19**:767-774.
34. Chatelain F, Tournieret JY, Ferrari A, Inglada J: Bivariate gamma distributions for image registration and change detection. *IEEE Trans Image Process* 2007; **17**:1796-1806.

35. White IR, Walker S, Babiker A. strbee: Randomization-based efficacy estimator. *STATA Journal* 2002; **2**:140-150.
36. Michael JR, Schucany WR. The mixture approach for simulating bivariate distributions with specified correlations. *American Statistician* 2002; **56**:48-54.
37. Gutierrez RG, Carter SL, Drukker DM: On boundary-value likelihood-ratio tests. *Stata Technical Bulletin* 2001; **60**:15-18.
38. Claeskens G, Nguti R, Janssen P. One-sided tests in shared frailty models. *Test* 2017; **17**:69-82.
39. McNamee R: Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Stat Med* 2009; **28**:2639-2652.
40. Odondi L, McNamee R. Performance of statistical methods for analysing survival data in the presence of non-random compliance. *Stat Med* 2010; **29**:2994-3003.
41. Goetghebeur E, Molenberghs G, Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Stat Med* 1998; **17**:341-355.
42. Pocock SJ, Abdalla M. The hope and the hazards of using compliance data in randomized controlled trials. *Stat Med* 1998; **17**:303-317.
43. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. *Lancet* 2013; **381**:752-762.
44. Moolgavkar SH. Commentary: Frailty and heterogeneity in epidemiological studies. *Int J Epidemiol* 2015; **44**:1425-1426.
45. Furman E. On a multivariate gamma distribution. *Statistics and Probability Letters* 2008; **78**:2353-2360.

Table 1: Simulation median event time and $H(t_{\text{med}})$ by $\exp(\beta^E)$, κ and type of censoring

(a) Mean (mths) across simulations of the trial median event time

$\exp(\beta^E)$	Weibull shape, κ	Simple Censoring	Inf censoring $\exp(\beta^C)=0.5$	Inf censoring $\exp(\beta^C)=1$	Inf censoring $\exp(\beta^C)=2$
0.50	0.75	7.2	6.0	5.7	5.3
	1	9.7	8.5	8.1	7.8
	1.50	13.1	12.0	11.7	11.3
0.75	0.75	7.0	5.9	5.6	5.2
	1	9.5	8.3	8.0	7.6
	1.50	12.9	11.8	11.5	11.1
1.00	0.75	6.7	5.7	5.4	5.0
	1	9.2	8.1	7.8	7.4
	1.50	12.7	11.7	11.3	10.9
1.50	0.75	6.2	5.3	5.0	4.6
	1	8.7	7.7	7.4	7.0
	1.50	12.2	11.2	10.9	10.5
2.00	0.75	5.7	4.8	4.6	4.3
	1	8.2	7.2	6.9	6.6
	1.5	11.7	10.8	10.5	10.1

(b) Control group cumulative hazard at the trial median event time*

0.50	0.75	0.2420	0.2116	0.2039	0.1944
	1	0.2418	0.2114	0.2037	0.1941
	1.50	0.2417	0.2112	0.2034	0.1939
0.75	0.75	0.2376	0.2085	0.2004	0.1899
	1	0.2375	0.2083	0.2001	0.1897
	1.50	0.2373	0.2081	0.1999	0.1895
1.00	0.75	0.2310	0.2036	0.1955	0.1853
	1	0.2309	0.2035	0.1953	0.1851
	1.50	0.2307	0.2033	0.1951	0.1849
1.50	0.75	0.2172	0.1920	0.1845	0.1747
	1	0.2171	0.1919	0.1844	0.1745
	1.50	0.2169	0.1917	0.1842	0.1743
2.00	0.75	0.2046	0.1807	0.1737	0.1643
	1	0.2045	0.1806	0.1736	0.1642
	1.50	0.2044	0.1804	0.1734	0.1640

* evaluated by $H(t) = \lambda t^\kappa$ with t = mean of the median event times in each simulation set

Table 2: Estimation of $\exp(\beta^E)$ using smPH method and AFT Weibull HOM methods when there is UH, and with and without informative censoring. Simulation summary measures* are $\exp(\hat{\beta}^E)$ and SB% bias in β^E . Predictions from (10-11) summarised similarly. Results in bold are those for which t-test gave $p > 0.05$.

$\exp(\beta^E)$	Prediction/ Analysis Method	Simple Censoring	Inf. censoring $\exp(\beta^C)=0.5$	Inf. censoring $\exp(\beta^C)=1$	Inf. censoring $\exp(\beta^C)=2.0$
0.50	Prediction	0.553; -14.4%	0.564; -17.4%	0.542; -11.5%	0.508; -2.3%
	Cox PH	0.553; -14.7%	0.564; -17.5%	0.542; -11.6%	0.508; -2.4%
	AFT HOM	0.552; -14.4%	0.562; -16.9%	0.541; -11.2%	0.508; -2.4%
0.75	Prediction	0.787; -16.7%	0.809; -26.5%	0.779; -13.4%	0.735; +7.2%
	Cox PH	0.787; -16.7%	0.807; -25.3%	0.778; -12.7%	0.736; +6.4%
	AFT HOM	0.786; -16.3%	0.804; -24.3%	0.777; -12.3%	0.738; +5.8%
1.00	Prediction	1.000	1.035	1.000	0.947
	Cox PH	0.999 (p=0.66)	1.031	0.998 (p=0.58)	0.950
	AFT HOM	0.999 (p=0.68)	1.029	0.998 (p=0.57)	0.953
1.50	Prediction	1.380; -20.8%	1.442; -9.8%	1.401; -16.9%	1.336; -28.5%
	Cox PH	1.375; -21.5%	1.432; -11.4%	1.394; -18.1%	1.336; -28.5%
	AFT HOM	1.378; -20.9%	1.432; -11.4%	1.397; -17.5%	1.344; -27.2%
2.00	Prediction	1.716; -22.1%	1.805; -14.8%	1.762; -18.3%	1.692; -24.1%
	Cox PH	1.704; -23.1%	1.789; -16.1%	1.748; -19.4%	1.684; -24.8%
	AFT HOM	1.711; -22.5%	1.793; -15.8%	1.756; -18.8%	1.697; -23.7%

* Results did not vary with κ : see text

Table 3: Estimation of Acceleration parameter, β^{*E} and Weibull shape, κ , from an AFT Weibull HOM model, when there is UH, and with and without informative censoring.

Simulation summary measures are $\exp(-\hat{\beta}^{*E})$, SB% bias for β^{*E} , $\hat{\kappa}$ and SB% bias for κ .

Results in bold are those for which t-test gave $p > 0.05$.

$\exp(\beta^E)$	$\kappa : \exp(-\beta^{*E})$	Summary statistic	Simple Censoring	Inf. censoring $\exp(\beta^C)=0.5$	Inf. censoring $\exp(\beta^C)=1$	Inf. censoring $\exp(\beta^C)=2.0$
0.5	0.397 0.500 0.630	$\exp(-\hat{\beta}^{*E})$	0.419; -6.0%	0.422; -6.8%	0.396; 0.3%	0.358; 11.1%
			0.521; -6.0%	0.524; -6.8%	0.499; 0.3%	0.463; 11.1%
			0.648; -6.0%	0.650; -6.8%	0.629; 0.3% each p=0.73	0.599; 11.1%
	0.75 1.00 1.50	$\hat{\kappa}$	0.685; -8.7%	0.670; -10.6%	0.666; -11.2%	0.661; -11.8%
			0.913; -8.7%	0.894; -10.6%	0.888; -11.2%	0.882; -11.2%
			1.370; -8.7%	1.341; -10.6%	1.332; -11.2%	1.323; -11.2%
0.75	0.681 0.750 0.825	$\exp(-\hat{\beta}^{*E})$	0.701; -7.5%	0.721; -14.6%	0.682; -0.4%	0.629; 21.1%
			0.766; -7.5%	0.782; -14.6%	0.751; -0.4%	0.706; 21.1%
			0.837; -7.5%	0.849; -14.6%	0.826; -0.4% each p=0.82	0.793; 21.1%
	0.75 1.00 1.50	$\hat{\kappa}$	0.680; -9.3%	0.666; -11.2%	0.662; -11.7%	0.658; -12.3%
			0.907; -9.3%	0.888; -11.2%	0.883; -11.7%	0.877; -12.3%
			1.360; -9.3%	1.332; -11.2%	1.325; -11.7%	1.316; -12.3%
1.00	1.000 1.000 1.000	$\exp(-\hat{\beta}^{*E})$	0.998	1.044	0.997	0.928
			0.998	1.033	0.997	0.946
			0.999 each p=0.67	1.022	0.998 each p=0.56	0.964
	0.75 1.00 1.50	$\hat{\kappa}$	0.673; -10.2%	0.661; -11.9%	0.657; -12.4%	0.654; -12.9%
			0.898; -10.2%	0.881; -11.9%	0.876; -12.4%	0.871; -12.9%
			1.346; -10.2%	1.321; -11.9%	1.315; -12.4%	1.307; -12.9%
1.50	1.717 1.500 1.310	$\exp(-\hat{\beta}^{*E})$	1.627; -9.9%	1.742; 2.7%	1.679; -4.1%	1.583; -15.0%
			1.441; -9.9%	1.516; 2.7%	1.475; -4.1%	1.411; -15.0%
			1.276; -9.9%	1.320; 2.7%	1.296; -4.1%	1.258; -15.0%
	0.75 1.00 1.50	$\hat{\kappa}$	0.659; -12.1%	0.649; -13.5%	0.646; -13.8%	0.644; -14.2%
			0.879; -12.1%	0.865; -13.5%	0.862; -13.8%	0.858; -14.2%
			1.319; -12.1%	1.297; -13.5%	1.293; -13.8%	1.287; -14.2%
2.00	2.520 2.000 1.587	$\exp(-\hat{\beta}^{*E})$	2.298; -10.0%	2.503; -0.7%	2.427; -4.1%	2.307; -9.6%
			1.867; -10.0%	1.990; -0.7%	1.944; -4.1%	1.872; -9.6%
			1.516; -10.0%	1.582; -0.7% each p=0.21	1.558; -4.1%	1.519; -9.6%
	0.75 1.00 1.50	$\hat{\kappa}$	0.647; -13.8%	0.638; -15.0%	0.636; -15.2%	0.634; -15.5%
			0.862; -13.8%	0.850; -15.0%	0.848; -15.2%	0.845; -15.5%
			1.293; -13.8%	1.275; -15.0%	1.272; -15.2%	1.268; -15.5%

Table 4: Estimation of β^{E*} , using G estimation and AFT Weibull HET models, given UH and with and without informative censoring. Simulation summary measures are $\exp(-\hat{\beta}^{*E})$ and SB% bias for β^{*E} . Results in bold: $p > 0.05$ in a t-test with $\kappa = 0.75$.

$\exp(\beta^E)$	$\kappa, \exp(-\beta^{*E})$	Analysis Method	Simple Censoring	Inf. censoring $\exp(\beta^C)=0.5$	Inf. censoring $\exp(\beta^C)=1$	Inf. censoring $\exp(\beta^C)=2.0$
0.5	0.75, 0.397 1.00, 0.500 1.50, 0.630	G estim	0.394; 0.8% 0.497; 0.8% 0.628; 0.8% p=0.24	0.394; 0.9% 0.496; 0.9% 0.627; 0.9%	0.370; 7.7% 0.474; 7.7% 0.608; 7.7%	0.337; 17.8% 0.442; 17.8% 0.580; 17.8%
	0.75, 0.397 1.00, 0.500 1.50, 0.630	AFT HET	0.392; 1.3% 0.495; 1.3% 0.626; 1.3% p=0.05	0.392; 1.3% 0.496; 1.3% 0.626; 1.3%	0.369; 7.9% 0.473; 7.9% 0.607; 7.9%	0.336; 17.9% 0.442; 17.9% 0.580; 17.9%
0.75	0.75, 0.681 1.00, 0.750 1.50, 0.825	G estim	0.676; 1.9% 0.746; 1.9% 0.822; 1.9% p=0.21	0.696; -5.6% 0.762; -5.6% 0.834; -5.6%	0.657; 9.6% 0.730; 9.6% 0.810; 9.6%	0.602; 32.2% 0.684; 32.2% 0.776; 32.2%
	0.75, 0.681 1.00, 0.750 1.50, 0.825	AFT HET	0.677; 1.6% 0.746; 1.6% 0.823; 1.6% p=0.27	0.696; -5.4% 0.762; -5.4% 0.834; -5.4%	0.657; 9.3% 0.730; 9.3% 0.811; 9.3%	0.605; 31.0% 0.686; 31.0% 0.778; 31.0%
1.00	0.75, 1.000 1.00, 1.000 1.50, 1.000	G estim	0.997 0.998 0.999 p=0.60	1.051 1.038 1.025	0.996 0.997 0.998 p=0.55	0.918 0.938 0.958
	0.75, 1.000 1.00, 1.000 1.50, 1.000	AFT HET	0.997 0.998 0.999 p=0.62	1.046 1.034 1.023	0.996 0.997 0.998 p=0.53	0.923 0.942 0.961
1.50	0.75, 1.717 1.00, 1.500 1.50, 1.310	G estim	1.717; 0.0% 1.500; 0.0% 1.310; 0.0% p=1.00	1.848; 13.6% 1.585; 13.6% 1.359; 13.6%	1.776; 6.2% 1.538; 6.2% 1.332; 6.2%	1.661; -6.2% 1.463; -6.2% 1.289; -6.2%
	0.75, 1.717 1.00, 1.500 1.50, 1.310	AFT HET	1.715; -0.2% 1.499; -0.2% 1.310; -0.2% p=0.85	1.838; 12.6% 1.578; 12.6% 1.356; 12.6%	1.771; 5.7% 1.535; 5.7% 1.331; 5.7%	1.665; -5.7% 1.466; -5.7% 1.290; -5.7%
2.00	0.75, 2.520 1.00, 2.000 1.50, 1.587	G estim	2.520; 0.0% 2.000; 0.0% 1.587; 0.0% p=0.98	2.739; 9.0% 2.129; 9.0% 1.655; 9.0%	2.657; 5.7% 2.081; 5.7% 1.630; 5.7%	2.516; -0.2% 1.998; -0.2% 1.586; -0.2%
	0.75, 2.520 1.00, 2.000 1.50, 1.587	AFT HET	2.518; -0.1% 1.999; -0.1% 1.587; -0.1% p=0.89	2.727; 8.5% 2.122; 8.5% 1.651; 8.5%	2.649; 5.4% 2.076; 5.4% 1.627; 5.4%	2.520; 0.0% 2.000; 0.0% 1.588; 0.0%

Table 5: Bias in smPH method in relation to the censoring fraction under two forms of censoring and approximations for predicted bias: see [9] and equation (10)

	censoring fraction					
	0.20		0.50		0.80	
(a) Results for ‘random’ censoring as defined in reference ([9])						
β	β_{smPH}/β	H&O approx	β_{smPH}/β	H&O approx	β_{smPH}/β	H&O approx
1	0.59	0.59	0.71	0.71	0.84	0.84
2	0.59	0.59	0.70	0.71	0.83	0.84
(b) Results from simulations with fixed M and prediction from [10]						
β	95% CI for β_{smPH}/β	from (10)	95% CI for β_{smPH}/β	from (10)	95% CI for β_{smPH}/β	from (10)
-0.5	(0.538,0.546)	0.534	(0.707,0.717)	0.703	(0.873,0.889)	0.880
0.5	(0.540,0.547)	0.535	(0.708,0.718)	0.703	(0.878,0.894)	0.880
1	(0.537,0.541)	0.537	(0.708,0.713)	0.710	(0.884,0.893)	0.885
2	(0.521,0.523)	0.551	(0.700,0.703)	0.733	(0.897,0.902)	0.901
	$H(t_{med})$		$H(t_{med})$		$H(t_{med})$	
-0.5	0.908		0.437		0.141	
0.5	0.550		0.266		0.085	
1	0.423		0.197		0.062	
2	0.240		0.100		0.029	

Table A1: Observed $SD(\hat{\beta}^E)$ of smPH and AFT Weibull HOM methods in simulation sets*; results were identical for sets with $\kappa=0.75$, $\kappa=1.00$ and $\kappa=1.50$.

$exp(\beta^E)$	Analysis Method	Simple Censoring	Inf. censoring $exp(\beta^C)=0.5$	Inf. censoring $exp(\beta^C)=1$	Inf. censoring $exp(\beta^C)=2.0$
0.50	Cox PH AFT HOM	0.1198 0.1203	0.1335 0.1341	0.1382 0.1388	0.1457 0.1463
0.75	Cox PH AFT HOM	0.1120 0.1125	0.1232 0.1239	0.1278 0.1284	0.1346 0.1353
1.00	Cox PH AFT HOM	0.1063 0.1069	0.1162 0.1169	0.1202 0.1210	0.1265 0.1274
1.50	Cox PH AFT HOM	0.1000 0.1008	0.1081 0.1090	0.1126 0.1135	0.1173 0.1185
2.00	Cox PH AFT HOM	0.0971 0.0981	0.1058 0.1068	0.1085 0.1096	0.1128 0.1141

* Corresponding results for bias are shown in Table 2.

Table A2: Observed $SD(\hat{\beta}^E)$ of G estimation and AFT Weibull HET methods in simulation sets*; results shown separately for sets with $\kappa = 0.75$, $\kappa = 1.00$ and $\kappa = 1.50$.

$exp(\beta^E)$	$\kappa, exp(\beta^{*E})$	Analysis Method	Simple Censoring	Inf. censoring $exp(\beta^C)=0.5$	Inf. censoring $exp(\beta^C)=1$	Inf. censoring $exp(\beta^C)=2.0$
0.5	0.75, 0.397 1.00, 0.500 1.50, 0.630	G estim	0.2017 0.1520 0.1013	0.2260 0.1695 0.1130	0.2324 0.1743 0.1162	0.2434 0.1825 0.1217
	0.75, 0.397 1.00, 0.500 1.50, 0.630	AFT HET	0.1970 0.1477 0.0985	0.2222 0.1666 0.1111	0.2286 0.1714 0.1143	0.2417 0.1813 0.1208
0.75	0.75, 0.681 1.00, 0.750 1.50, 0.825	G estim	0.1852 0.1389 0.0926	0.2095 0.1571 0.1047	0.2170 0.1628 0.1085	0.2259 0.1694 0.1129
	0.75, 0.681 1.00, 0.750 1.50, 0.825	AFT HET	0.1822 0.1366 0.0911	0.2052 0.1539 0.1026	0.2133 0.1600 0.1067	0.2240 0.1680 0.1120
1.00	0.75, 1.000 1.00, 1.000 1.50, 1.000	G estim	0.1798 0.1348 0.0899	0.2008 0.1506 0.1004	0.2083 0.1563 0.1042	0.2196 0.1647 0.1098
	0.75, 1.000 1.00, 1.000 1.50, 1.000	AFT HET	0.1751 0.1313 0.0875	0.1950 0.1463 0.0975	0.2034 0.1526 0.101	0.2135 0.1601 0.1068
1.50	0.75, 1.717 1.00, 1.500 1.50, 1.310	G estim	0.1762 0.1322 0.0881	0.1906 0.1429 0.0953	0.1994 0.1495 0.0997	0.2097 0.1573 0.1049
	0.75, 1.717 1.00, 1.500 1.50, 1.310	AFT HET	0.1717 0.1288 0.0859	0.1868 0.1401 0.0934	0.1952 0.1464 0.0976	0.2041 0.1531 0.1021
2.00	0.75, 2.520 1.00, 2.000 1.50, 1.587	G estim	0.1759 0.1319 0.0880	0.1899 0.1424 0.0949	0.1958 0.1469 0.0979	0.2044 0.1533 0.1022
	0.75, 2.520 1.00, 2.000 1.50, 1.587	AFT HET	0.1715 0.1286 0.0858	0.1853 0.1390 0.0927	0.1905 0.1429 0.0953	0.1995 0.1496 0.0997

* Corresponding results for bias are shown in Table 4.