Research Article

Statistics in Medicine

Received 5 October 2011, Accepted 6 January 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5327

An efficient alternative to the stratified Cox model analysis

Devan V. Mehrotra,[a*†] Shu-Chih Su[a] and Xiaoming Li[b]

Consider a typical two-treatment randomized clinical trial involving a time-to-event endpoint, with randomization stratified by a categorical prognostic factor (for example gender). At the design stage, it is often assumed that the treatment hazard ratio (HR) is constant across the strata, and the data are commonly analyzed using the stratified Cox proportional hazards model. We caution that this ubiquitous approach is needlessly risky because departures from the assumption of the HR being the same for all the strata can result in a notably biased and/or less powerful analysis. An alternative approach is proposed in which first the [log] HR is estimated separately for each stratum using an unstratified Cox model, and then the stratum-specific estimates are combined for overall inference using either sample size or 'minimum risk' stratum weights. The advantages of the proposed two-step analysis versus the common one-step stratified Cox model analysis are illustrated using simulations that were conducted to support the design of a vaccine clinical trial. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: inverse variance weights; minimum risk weights; sample size weights; stratified logrank; treatment by stratum interaction

1. Introduction

For a typical randomized clinical trial with a time-to-event endpoint, if the risk of experiencing the event of interest (for example myocardial infarction) during the planned duration of the trial is expected to be systematically influenced by one or more prognostic factors (for example gender), then the strategy of prestratification is often employed. First, each subject is mapped to a stratum based on his/her prognostic profile. Then, within each stratum, the subjects are randomized to either treatment A or treatment B, and their 'survival' times, that is times from randomization to the occurrence of the event of interest, are recorded. For example, during the design stage of a placebo-controlled clinical trial of an experimental vaccine that motivated this research, it was recognized up front that the overall trial population was an approximately equal mixture of two subgroups defined by a binary baseline covariate, and it was anticipated that the event rate in the placebo arm would be lower in the first subgroup relative to the second. Accordingly, the plan was to stratify randomization, with stratum 1 and stratum 2 representing the subgroups with lower and higher baseline event risk, respectively.

For the remainder of this article, we focus on typical settings in which the trial enrollment is stratified on the basis of (say) one to three categorical prognostic factor(s); in most applications, the resulting total number of strata is small, typically two to eight. We assume throughout that, within each stratum, the two survival time distributions have proportional hazard functions. In other words, in stratum i, $\theta_i = \lambda_{i\mathrm{A}}(t) \div \lambda_{i\mathrm{B}}(t)$ is constant over time, where $\lambda_{i\mathrm{A}}(t)$ and $\lambda_{i\mathrm{B}}(t)$ denote the hazard function at time t for treatments A and B in stratum i, respectively. The presumed primary objective of the trial is to estimate the *overall* hazard ratio parameter, denoted by θ and defined below, and to test the null hypothesis $H_0 : \theta = \theta_0$ {or equivalently $H_0 : \beta = \beta_0$, where $\beta = \ln(\theta)$} versus the one-tailed alternative $H_1 : \theta < \theta_0$; a two-tailed alternative can be readily accommodated. The ubiquitous tool for estimation and inference in this setting is the stratified version of the Cox proportional hazards model [1].

[a]*Merck Research Laboratories, 351 N. Sumneytown Pike, North Wales, PA 19454, USA*
[b]*Gilead Sciences, 199 East Blaine Street, Seattle, WA 98102, USA*
**Correspondence to: Devan V. Mehrotra, Clinical Biostatistics, UG1CD-44, Merck Research Laboratories 351 N. Sumneytown Pike, North Wales, PA 19454, USA.*
[†]*E-mail: devan_mehrotra@merck.com*

The stratified Cox model analysis implicitly assumes that the treatment hazard ratio is constant across strata. If this assumption is incorrect, then use of the resulting misspecified model can result in a biased and/or less powerful analysis; to be clear, the term *bias* here is with respect to θ being defined as the antilogarithm of the population-weighted average of the stratum-specific log hazard ratios, as detailed in Section 2. To circumvent the risk of bias and associated consequences, we propose an alternative approach that does not require the assumption of a common hazard ratio across strata. We describe the stratified Cox model analysis and our proposed alternative in the next section. In Section 3, we contrast the different analytic approaches using simulations, and we conclude with summary remarks in Section 4.

2. Stratified Cox model analysis and proposed alternative

The common stratified Cox proportional hazards model accommodates distinct baseline hazard functions $(\lambda_{i0}(t) \quad i = 1 \ldots s)$ for each stratum, but enforces a common treatment hazard ratio across strata. For example, if the subjects within a stratum are homogeneous (which is what we assume throughout for simplicity), the hazard function at time t for a subject in stratum i is $\lambda_i(t|Z) = \lambda_{i0}(t)\, e^{Z\beta}$,where $Z = 1$ for treatment A and 0 for treatment B, and $\beta = \ln(\theta)$; $\theta = e^{\beta}$ is the hazard ratio. Suppose there are k_i observed events in stratum i, with distinct ordered event times $t_{i1} < t_{i2} < \ldots < t_{ik_i}$. If there are n_{ijA} and n_{ijB} subjects at risk just before time t_{ij} in treatment groups A and B, respectively, and the binary indicator d_{ijA} equals 1 (0) if the event at time t_{ij} in stratum i is from treatment A (B), then the Cox partial likelihood function is

$$L(\beta) = \prod_{i=1}^{s} \prod_{j=1}^{k_i} \frac{e^{d_{ij\mathrm{A}}\beta}}{n_{ij\mathrm{A}} e^{\beta} + n_{ij\mathrm{B}}}. \tag{1}$$

The score function, $S(\beta) = \partial \ln L(\beta) / \partial\beta$, arising from (1) is

$$S(\beta) = \sum_{i=1}^{s} \sum_{j=1}^{k_i} \left(d_{ijA} - \frac{n_{ij\mathrm{A}} e^{\beta}}{n_{ij\mathrm{A}} e^{\beta} + n_{ij\mathrm{B}}} \right). \tag{2}$$

The Cox maximum partial likelihood estimator of β denoted by $\hat{\beta}$, obtained as the iterative solution of $S(\beta) = 0$, has an asymptotic normal distribution [2, 3] with mean β and approximate variance $V\left(\hat{\beta}\right) = I(\beta)^{-1}$, where

$$I(\beta) = -\frac{\partial^2 \ln L(\beta)}{\partial\beta^2} = \sum_{i=1}^{s} \sum_{j=1}^{k_i} \frac{n_{ij\mathrm{A}} n_{ij\mathrm{B}} e^{\beta}}{\left(n_{ij\mathrm{A}} e^{\beta} + n_{ij\mathrm{B}}\right)^2}. \tag{3}$$

The corresponding stratified Cox model estimator of θ is $\hat{\theta} = e^{\hat{\beta}}$. In addition to obtaining a point estimate of β (or of $\theta = e^{\beta}$), interest lies in testing $H_0 : \beta = \beta_0$ (or $H_0 : \theta = \theta_0$) and providing a confidence interval (CI) for β. This is commonly conducted using either the Wald or score approach; the former is more popular presumably because of its availability in commercial software. For the Wald approach, the test statistic is $Z(\beta_0) = \left(\hat{\beta} - \beta_0\right) / \sqrt{\hat{V}\left(\hat{\beta}\right)}$, and an approximate $100(1-\alpha)\,\%$ confidence interval for β is $\hat{\beta} \pm Z_{\alpha/2} \sqrt{\hat{V}\left(\hat{\beta}\right)}$, where $\hat{V}\left(\hat{\beta}\right)$ is obtained by replacing β with $\hat{\beta}$ in (3) and taking the reciprocal. The test statistic based on the score approach is given by

$$U(\beta_0) = \frac{\{S(\beta_0)\}^2}{I(\beta_0)}. \tag{4}$$

Provided the number of events is sufficiently large, $U(\beta_0)$ is approximately distributed as a $\chi^2_{(1)}$ variate under the null hypothesis. Of note, when $\beta_0 = 0$ (that is $\theta_0 = 1$), (4) leads to the well-known stratified version of the logrank statistic [4].

If there is no treatment by stratum interaction, that is if $\theta_i = \theta$ for all i, there is no ambiguity about the definition of the overall hazard ratio θ. However, in the presence of an interaction, that is if $\theta_i \neq \theta_{i*}$ for

at least one i and $i*$, there are at least two ways to define θ: a population weighted average of the θ_i's, or the antilogarithm of a population weighted average of the β_i's, where $\beta_i = \ln(\theta_i)$. Because estimation for the Cox model is accomplished via log hazard ratios, it is reasonable to use the latter definition so that $\theta = e^{\bar{\beta}}$, where

$$\bar{\beta} = \sum_{i=1}^{s} f_i \beta_i. \tag{5}$$

In (5), f_i is the fraction of subjects in the target population that are from stratum i ($\sum_{i=1}^{S} f_i = 1$), that is if all the subjects in the population to which the trial results are going to be generalized were (hypothetically) enrolled, f_i is the resulting fraction that would be from stratum i. If a total of N subjects are enrolled in the clinical trial, of which N_i are from stratum i, we assume hereafter that $N_i/N \approx f_i$, which is reasonable for a sufficiently large trial It is important to acknowledge that it is difficult to interpret $\bar{\beta}$ in (5) if the individual β_i's vary 'considerably'. We return to this issue in Section 3, and for the remainder of this article we make the assumption that there is no *qualitative* treatment by stratum interaction, that is we assume either $\beta_i \geqslant \beta_0$ or $\beta_i \leqslant \beta_0$ for all i We stress that this assumption is for interpretational purposes only; it is not required for the validity of the proposed alternative to the one-step stratified Cox model analysis discussed next.

To make allowance for a true treatment by stratum interaction, the relevant partial likelihood function is

$$L(\beta_1, \ldots, \beta_s) = \prod_{i=1}^{s} \prod_{j=1}^{k_i} \frac{e^{d_{ij\mathrm{A}}\beta_i}}{n_{ij\mathrm{A}} e^{\beta_i} + n_{ij\mathrm{B}}}. \tag{6}$$

If (6) is correct, as explained later, $\hat{\beta}$ obtained via maximization of (1) will generally represent a biased estimator of $\bar{\beta}$, and hence of the overall hazard ratio $\theta = e^{\bar{\beta}}$. Accordingly, a natural alternative to the common stratified Cox model analysis described above is to consider estimators of the form

$$\hat{\bar{\beta}}_w = \sum_{i=1}^{s} w_i \hat{\beta}_i \tag{7}$$

where w_i is the weight assigned to stratum i ($\sum_{i=1}^{s} w_i = 1$), and $\hat{\beta}_i$ is the Cox maximum partial likelihood estimator of β_i with estimated variance $\hat{V}\left(\hat{\beta}_i\right) \equiv \hat{V}_i$. An approximate Wald-type confidence interval for $\bar{\beta}$ is $\hat{\bar{\beta}}_w \pm Z_{\alpha/2}\sqrt{\hat{V}\left(\hat{\bar{\beta}}_w\right)}$, with $\hat{V}\left(\hat{\bar{\beta}}_w\right) = \sum_{i=1}^{s} w_i^2 \hat{V}\left(\hat{\beta}_i\right)$. Note that in contrast to the one-step stratified Cox model analysis involving direct maximization of (1), the proposed alternative analysis involves two steps: estimation of the individual β_i's in step 1 followed by a synthesis of the stratum-specific estimates via (7) in step 2.

We consider two weighting options in (7). The first entails using 'sample size (SSIZE)' weights, $w_i^{\mathrm{SSIZE}} = N_i/N$, motivated by a desire to obtain an asymptotically unbiased estimator of $\bar{\beta}$ regardless of whether or not the β_i's are the same. However, the potential price of seeking unbiasedness is that the resulting estimator may have a needlessly large variance. Accordingly, the second weighting option is to use an analogue of the 'minimum risk (MR)' weights proposed by Mehrotra and Railkar [5] for stratified comparative binomial trials The MR weighting strategy is intended to minimize the mean squared error (MSE = squared bias + variance) when estimating $\bar{\beta}$. The weights can be easily calculated using the plug-in formula

$$w_i^{\mathrm{MR}} = \frac{d_i}{\sum_{i=1}^{s} \hat{V}_i^{-1}} - \left(\frac{c_i \hat{V}_i^{-1}}{\sum_{i=1}^{s} \hat{V}_i^{-1} + \sum_{i=1}^{s} c_i \hat{\beta}_i \hat{V}_i^{-1}}\right)\left(\frac{\sum_{i=1}^{s} \hat{\beta}_i d_i}{\sum_{i=1}^{s} \hat{V}_i^{-1}}\right) \tag{8}$$

where $c_i = \hat{\beta}_i \sum_{i=1}^{s} \hat{V}_i^{-1} - \sum_{i=1}^{s} \hat{\beta}_i \hat{V}_i^{-1}$ and $d_i = \hat{V}_i^{-1}\left(1 + c_i \sum_{i=1}^{s} (N_i/N)\hat{\beta}_i\right)$. Note that when $\hat{\beta}_i \approx$constant, $w_i^{\mathrm{MR}} \approx \hat{V}_i^{-1} / \sum_{i=1}^{s} \hat{V}_i^{-1}$, an appealing result because inverse-variance (INVAR)

weighting is asymptotically optimal from a MSEminimizing perspective if the β_i's are indeed the same. Building on this observation, note that we could have considered yet another weighting option in (7), namely $w_i^{\text{INVAR}} = \hat{V}_i^{-1} / \sum_{i=1}^{s} \hat{V}_i^{-1}$. This would be a reasonable choice of weights if our singular goal was to try to minimize the variance of the two-step estimator in large samples, regardless of bias. However, if the β_i's are not the same, using INVAR weights will generally yield a biased estimate of $\bar{\beta}$, and the magnitude of the bias can be nontrivial in some cases. A rough approximation of the large-sample bias can be determined as follows. With r:1 (A:B) randomization, and for small to moderate values of $|\beta_i|$ that are typically expected in practice, because most $n_{ij\text{A}}/n_{ij\text{B}} \approx r$ when the sample sizes (or more accurately, the risk sets) are large, it follows from (3) that $I(\beta_i) = V_i^{-1} \approx sk_i re^{\beta_i} \left(re^{\beta_i}+1\right)^{-2}$. Because $\hat{\beta}_i \approx \beta_i$ asymptotically, and using the approximation that $E\left(k_i / \sum_{i=1}^{s} k_i\right) \approx E(k_i) / E\left(\sum_{i=1}^{s} k_i\right)$, it follows that the approximate magnitude of the large-sample bias associated with $\hat{\bar{\beta}}_{\text{INVAR}} = \sum_{i=1}^{s} w_i^{\text{INVAR}} \hat{\beta}_i$ is given by

$$E\left[Bias\left(\hat{\bar{\beta}}_{\text{INVAR}}\right)\right] = E\left(\hat{\bar{\beta}}_{\text{INVAR}} - \bar{\beta}\right) \approx \sum_{i=1}^{s} (a_i - f_i)\,\beta_i \tag{9}$$

where $a_i = E(k_i)e^{\beta_i}(re^{\beta_i}+1)^{-2}\left[\sum_{i=1}^{s} E(k_i)e^{\beta_i}(re^{\beta_i}+1)^{-2}\right]^{-1}$.

If the β_i's are the same, that is if $\beta_i = \beta$ for all i, then the expected large-sample bias in (9) is zero; this follows because $\sum_{i=1}^{s} a_i\beta_i = \beta \sum_{i=1}^{s} E(k_i)\left[\sum_{i=1}^{s} E(k_i)\right]^{-1} = \beta$ and $\sum_{i=1}^{s} f_i\beta_i = \beta \sum_{i=1}^{s} f_i = \beta$, implying that $\sum_{i=1}^{s} (a_i - f_i)\,\beta_i = 0$. However, in other cases, the expected bias will either be negative or positive. Recall that we had stated earlier that the one-step stratified Cox model estimator can suffer from bias when the β_i's are not the same; this can now be explained heuristically by the link that the one-step estimator is conceptually similar (but not identical) to the two-step estimator with INVAR weights. In essence, (9) can be used to approximate the large-sample expected bias for the one-step estimator. The simulation results discussed in the next section support this observation.

3. Simulations to compare the one-step and two-step approaches

3.1. Simulation set-up

We conducted extensive simulations to compare the operating characteristics of the one-step and two-step approaches. The simulation set-up was motivated by the placebo-controlled trial of an experimental vaccine, as alluded to in Section 1; the vaccine is intended to help reduce the risk of acquiring a certain type of symptomatic infectious disease (details omitted). For vaccine efficacy trials, because the vaccine could ultimately be administered to millions of people, the typical regulatory requirement for licensure is to provide evidence of 'super efficacy' [6], that is to demonstrate with high confidence that the overall true vaccine efficacy, $VE = (1-\theta)$, is greater than a prespecified clinically meaningful threshold VE^*. In our simulations, we used $VE^* = 25\%$, implying a test of $H_0 : VE = 25\%$ versus $H_1 : VE > 25\%$; note that the true VE needs to be notably greater than 25% to deliver adequate power in a feasibly sized trial.

It was unknown at the design stage whether the vaccine efficacy would be the same or different for the two strata. Accordingly, to accommodate a potential treatment by stratum interaction, as noted in Section 2, the overall hazard ratio (vaccine: placebo) was defined as $\theta = e^{\bar{\beta}}$, with $\bar{\beta} = \sum_{i=1}^{2} f_i\beta_i$ as before. Because $\bar{\beta} = \ln(1-VE)$, the hypotheses of interest could be re-expressed as

$$H_0 : \bar{\beta} = -0.288 \text{ versus } H_1 : \bar{\beta} < -0.288. \tag{10}$$

Vaccine efficacy trials are often event-driven, that is follow-up of the randomized subjects continues until a prespecified target number of events have accrued during the trial. For the aforementioned vaccine trial, it was determined using the methodology described in Lachin [7] that 274 events would provide $\sim 90\%$ power to reject the null hypothesis in (10) at the one-tailed 2.5% alpha level based on the common stratified Cox model analysis assuming the true vaccine efficacy in each stratum was 50%, that is assuming $\beta_1 = \beta_2 = -0.693 \equiv \ln(1\text{–}0.5)$. To accrue 274 events in a reasonable time frame ($\sim$ 5 years) it was further determined that $\sim$ 6000 subjects would need to be randomized in total (1 : 1 randomization to vaccine : placebo) under certain assumptions about stratum-specific incident rates in the placebo group, rates of dropout (premature discontinuations from the trial) and so on.

Statistics in Medicine

We simulated the event-driven trial 5000 times, and the dataset from each simulated trial was subjected to the one-step stratified Cox model analysis, and the proposed two-step analyses with SSIZE and MR weighting; for completeness, INVAR weighting was also used to demonstrate the anticipated similarity between the one-step estimator and the two-step estimator with INVAR weights. Each such analysis produced a point estimate ($\bar{\beta}_{\text{est}}$) and corresponding 95% Wald CI for $\bar{\beta}$ (with $\bar{\beta} \equiv \beta$ in (1) for the one-step analysis), and the null hypothesis in (10) was rejected if the upper bound of the 95% CI was less than –0.288. For each analytic approach, we summarized (i) percent bias = $100 \times E\left(\bar{\beta}_{\text{est}} - \bar{\beta}\right) / \left|\bar{\beta}\right|$, (ii) MSE = $E\left(\bar{\beta}_{\text{est}} - \bar{\beta}\right)^2$, and (iii) proportion of times that the null hypothesis was rejected.

Five simulation scenarios were studied, as shown in Table I. (The overall log hazard ratio in Table I is based on $f_1 = f_2 = 0.5$; we did additional simulations with unequal stratum relative frequencies, namely $f_1 = 0.7$, $f_2 = 0.3$ and $f_1 = 0.3$, $f_2 = 0.7$, some results for which are also reported later.) Under the null hypothesis scenario, the vaccine efficacy was 25% in both strata (that is $\beta_1 = \beta_2 = -0.288 \equiv \ln(1 - .25)$), so the overall vaccine efficacy was 25%. Under all the four alternative hypothesis scenarios the overall vaccine efficacy was 50% with equal efficacy across the two strata (no interaction) in scenario 2, but unequal efficacies (interaction) in scenarios 3–5. Two important points are worth noting here. First, in all the interaction scenarios studied, the vaccine efficacy was unambiguously clinically meaningful (> 32%) in both strata, justifying the combination of stratum-specific efficacies into an 'overall' efficacy; such a combination would admittedly be hard to justify if one of the strata had less than clinically meaningful efficacy. Second, we chose the magnitudes of the true interactions, such that, as is often the case in practice, it would be relatively hard to detect a potential true interaction in a typical simulated trial. Specifically, the stratum-specific vaccine efficacies shown in Table I were chosen such that the power to detect a treatment by stratum interaction (at α=5%) in scenarios 3, 4, and 5 was approximately 20%, 40%, and 60%, respectively. With this set-up, we considered the magnitudes of the simulated interactions to be in the mild-to-moderate range.

3.2. *Simulation results*

For the five scenarios shown in Table I (with $f_1 = f_2 = 0.5$), Table II displays the estimated percent bias [top panel] and MSE-based relative efficiency [bottom panel] for each analytic approach, the latter calculated as 100 × MSE(stratified Cox model)/MSE(given approach). As expected, when $\beta_1 = \beta_2$ (no interaction scenarios 1 and 2), the one-step method yielded unbiased results, but as the difference between β_1 and β_2 increased, so did the magnitude of the bias, with the percent bias being notably large in scenarios 4 (9.7%) and 5 (13.4%); all the percent bias values for the one-step approach were close to those predicted by the approximate formula in (9). In contrast, by construction, the two-step analysis with SSIZE weights was unbiased in all scenarios. The two-step analysis with MR weights also performed well in this regard, with a narrow estimated percent bias range of –0.6% to 2.3% in the scenarios studied.

The results in the bottom panel of Table II show that, relative to the one-step analysis, the MSE based on the two-step analysis with either the SSIZE or MR weights was only slightly larger in scenarios 1–3,

Table I. Simulated scenarios.

#	Scenario	Stratum 1 β_1 $VE_1 = 100 \times \left(1 - e^{\beta_1}\right)$	Stratum 2 β_2 $VE_2 = 100 \times \left(1 - e^{\beta_2}\right)$	Overall* $\bar{\beta} = f_1\beta_1 + f_2\beta_2$ $VE = 100 \times \left(1 - e^{\bar{\beta}}\right)$
1	Null	–0.288	–0.288	–0.288
	(no interaction)	25%	25%	25%
2	Alt 1	–0.693	–0.693	–0.693
	(no interaction)	50%	50%	50%
3	Alt 2	–0.844	–0.541	–0.693
	(interaction)	57%	41.8%	50%
4	Alt 3	–0.916	–0.470	–0.693
	(interaction)	60%	37.5%	50%
5	Alt 4	–0.994	–0.392	–0.693
	(interaction)	63%	32.4%	50%

* $f_1 = f_2 = 0.5$

Table II. Simulation results.

#	Scenario	One-step analysis (stratified Cox model)	Two-step analysis (SSIZE weights)	Two-step analysis (MR weights)
Percent bias				
1	Null	–0.3	–0.7	–0.3
2	Alt 1	–0.5	–0.9	–0.6
3	Alt 2	4.9	–1.2	0.7
4	Alt 3	9.7	–0.1	2.3
5	Alt 4	13.4	–1.3	1.3
*Relative efficiency (%)**				
1	Null	100	95	98
2	Alt 1	100	95	98
3	Alt 2	100	96	99
4	Alt 3	100	111	112
5	Alt 4	100	123	125

*% RE = 100 × MSE(1-step analysis)/MSE(given method); $f_1 = f_2 = 0.5$; 5000 simulations.

Table III. Simulation results: % of times that the null hypothesis was rejected (α=2.5% one-tailed).*

#	Scenario	One-step analysis (stratified Cox model)	Two-step analysis (SSIZE weights)	Two-step analysis (MR weights)
1	Null	2.4	2.5	2.4
2	Alt 1	90	90	90
3	Alt 2	84	88	87
4	Alt 3	76	87	85
5	Alt 4	69	87	85

* $f_1 = f_2 = 0.5$; 5000 simulations.

but was notably smaller in the other scenarios; the latter result was largely driven by the bias of the one-step analysis. Of note, by construction, the MSE-based relative efficiency using the MR weights was numerically larger than that for the SSIZE weights.

For each method and under each of the five scenarios, Table III shows the proportion of times that the upper bound of the 95% CI for $\bar{\beta}$ was less than $\beta_0 = -0.288$ (null scenario: type I error rate; alternative scenarios: power). As expected, the type I error rate was controlled at the nominal one-tailed $\alpha = 2.5\%$ level for all the methods. Moreover, by design, when the vaccine efficacy was the same for the two strata, the power for the one-step analysis was 90%. However, the corresponding power was lower in the presence of a (relatively hard to detect) treatment by stratum interaction, strikingly so in scenarios 4 and 5, with powers of only 76% and 69%, respectively. Note that even though the overall vaccine efficacy was 50% in all the alternative hypothesis scenarios shown in Table I, the power of the one-step analysis ranged widely from 69% to 90% depending on how different the stratum-specific vaccine efficacies were. In contrast, use of the two-step analysis with either the SSIZE or the MR weights yielded 90% power in the no interaction scenario, and displayed only a small power loss when there was an interaction, with worst case powers of 87% and 85% for SSIZE and MR weights, respectively. This shows that the two-step analysis provides robustness against departures from the common no interaction assumption that can substantially weaken the one-step stratified Cox model analysis even in cases where the true vaccine efficacy is clinically meaningful in both strata.

The simulation results discussed above focused on the case $f_1 = f_2 = 0.5$; a similar pattern of results was observed for simulations with unequal stratum relative frequencies. A representative snapshot of those additional simulation results is provided in Figure 1 for the case $f_1 = 0.7$, $f_2 = 0.3$. The figure shows the empirical densities of the one-step and two-step estimators of $\bar{\beta}$, the latter using either SSIZE, MR, or INVAR weights; a vertical reference line is drawn in each panel at the true value of $\bar{\beta}$ to visually help assess potential bias; unlike the $f_1 = f_2 = 0.5$ case, here the true $\bar{\beta}$ values are slightly different in the Alt 1–Alt 4 scenarios. As expected, the one-step estimator and the two-step estimator with INVAR weights had very similar empirical densities; both estimators suffered from bias when $\beta_1 \neq \beta_2$, with bias

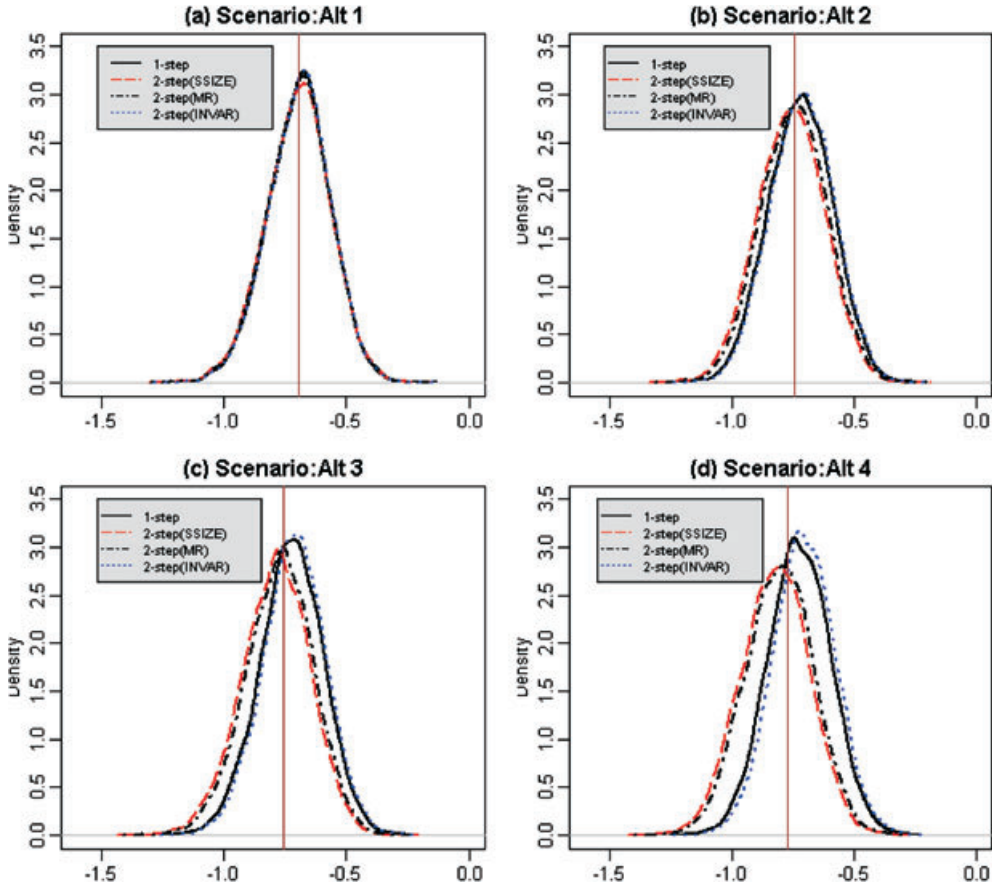

Figure 1. Empirical densities of the 1-step and 2-step estimator (5000 simulations) for $\overline{\beta} = f_1\beta_1 + f_2\beta_2$, with $f_1 = 0.7$, $f_2 = 0.3$. True parameters are $\beta_1 = -0.693$, $\beta_2 = -0.693$ in (a), $\beta_1 = -0.844$, $\beta_2 = -0.541$ in (b), $\beta_1 = -0.916$, $\beta_2 = -0.470$ in (c), and $\beta_1 = -0.994$, $\beta_2 = -0.392$ in (d). In each panel, the vertical reference line is at $\overline{\beta}$.

increasing as the distance between β_1 and β_2 increased. Importantly, as in Table II, the bias was positive in the interaction scenarios, thereby lowering statistical power. In contrast, the two-step estimators with SSIZE and MR weights were unbiased and minimally biased ($< 2\%$ bias), respectively. Both delivered similar power to that of the one-step approach when $\beta_1 = \beta_2$, but were notably more powerful when $\beta_1 \neq \beta_2$ (detailed results not shown for brevity, but available upon request).

4. Conclusions

For stratified time-to-event trials, it is common to assume at the design stage that the treatment effect, typically quantified using the (log) hazard ratio, is the same for all the strata, and to subsequently analyze the data using the stratified Cox model for estimation and inference. However, we have cautioned that the assumption of the treatment effect being exactly the same across strata is rarely justified and, indeed, unnecessary. Some allowance needs to be made for the reality that treatment effects may truly differ across strata. The common stratified Cox model approach, which implicitly uses a one-step analysis to estimate the 'overall' treatment effect, performs well in terms of bias and power if there is no treatment by stratum interaction. However, in the presence of a true (and often undetected) interaction, the estimator delivered by the one-step analysis can be potentially misleading. If the true overall log hazard ratio is defined as a weighted average of the true stratum-specific log hazard ratios, with weights corresponding to the true stratum relative frequencies in the population of interest, then it is clear that the stratified Cox model analysis will generally yield a biased estimate of this estimand. In our simulations for the vaccine trial, we showed that the resulting bias associated with even a modest treatment by stratum interaction can lead to a material loss in power, even when the vaccine efficacy is clinically meaningful in both strata.

Our proposed two-step analysis is a reasonable alternative to the usual stratified Cox model analysis. By first estimating the stratum-specific log hazard ratios, and then combining the results using either SSIZE or MR weighting, the two-step analysis delivers an estimator with negligible bias, if any. Moreover, it delivers power that is comparable to that for the one-step analysis when there is no treatment by stratum interaction, but notably better power when there is an interaction that adversely impacts the one-step estimator. It is worth reinforcing the latter point in practical terms. For the vaccine trial, among the scenarios shown in Table I, our simulations revealed that for the two-step analysis, the worst case power was 85% (MR weighting, scenario 5 in Table III). If, for some reason, there was a continued need to employ the usual one-step stratified Cox model analysis, to guarantee at least 85% power under every scenario studied, the target number of events would need to be increased by over 40%. This would require (needlessly) increasing the trial budget by millions of dollars!

While our research was motivated by a vaccine efficacy trial, it is important to note that the proposed two-step analysis is more broadly applicable for the analysis of stratified time-to-event data in other therapeutic areas. Moreover, while our simulations were confined to the case of two strata, there is no theoretical reason for the conclusions to change in other cases. In general, the number of strata should be kept small, with the choice of stratification factor(s) prudently driven by their prognostic value; overstratification can lead to a loss in power [8,9]. If the need to adjust for multiple stratification factors in the analysis results in 'small' sample sizes per treatment by stratum cell (say, less than 100 subjects), then the proposed two-step analysis can be potentially enhanced by replacing the Cox model estimator of the stratum-specific log hazard ratio with a corresponding estimator obtained by inverting the generalized logrank (GLR) test of Mehrotra and Roth [10]; the Cox model and GLR estimators are asymptotically similar, but the latter is notably more efficient in small samples. Research on the stratified extension of the GLR method is ongoing, and will be the subject of a separate article.

Acknowledgement

We thank two anonymous referees for helpful suggestions that led to an improved manuscript.

References

1. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
2. Tsiatis AA. A large sample study of Cox's regression model. *Annals of Statistics* 1981; **9**:93–108.
3. Anderson PK, Gill RD. Cox's regression model for counting processes: a large sample approach. *Annals of Statistics* 1982; **10**:1100–1120.
4. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; **50**:163–170.
5. Mehrotra DV, Railkar R. Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* 2000; **19**:811–825. DOI: 10.1002/(SICI)1097-0258(20000330).
6. Mehrotra DV. Vaccine clinical trials: a statistical primer. *Journal of Biopharmaceutical Statistics* 2006; **16**:403–414.
7. Lachin JM. *Biostatistical Methods: The Assessment of Relative Risks*. John Wiley & Sons, Inc.: New York, NY, 2000. 409–412.
8. Feng C, Wang H, Tu XM. Power loss of stratified log-rank test in homogeneous samples. *International Journal of Quality, Statistics, and Reliability* 2010:4. DOI: 10.1155/2010/942184.
9. De Stavola BL, Cox DR. On the consequences of overstratification. *Biometrika* 2008; **95**:992–996. DOI: 10.1093/biomet/asn039.
10. Mehrotra DV, Roth A. Relative risk estimation and inference using a generalized logrank statistic. *Statistics in Medicine* 2001; **20**:2099–2113. DOI: 10.1002/sim.854.