Received: 8 May 2018 | Revised: 22 October 2018 | Accepted: 20 December 2018

DOI: 10.1002/pst.1928

MAIN PAPER

WILEY

Hazard ratio inference in stratified clinical trials with time-to-event endpoints and limited sample size

Rengyi Xu[1] | Devan V. Mehrotra[2] | Pamela A. Shaw[1]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania

[2]Biostatistics and Research Decision Sciences, Merck Research Laboratories, North Wales, Pennsylvania

Correspondence
Rengyi Xu, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104.
Email: xurengyi@pennmedicine.upenn.edu

The stratified Cox model is commonly used for stratified clinical trials with time-to-event endpoints. The estimated log hazard ratio is approximately a weighted average of corresponding stratum-specific Cox model estimates using inverse-variance weights; the latter are optimal only under the (often implausible) assumption of a constant hazard ratio across strata. Focusing on trials with limited sample sizes (50-200 subjects per treatment), we propose an alternative approach in which stratum-specific estimates are obtained using a refined generalized logrank (RGLR) approach and then combined using either sample size or minimum risk weights for overall inference. Our proposal extends the work of Mehrotra et al, to incorporate the RGLR statistic, which outperforms the Cox model in the setting of proportional hazards and small samples. This work also entails development of a remarkably accurate plug-in formula for the variance of RGLR-based estimated log hazard ratios. We demonstrate using simulations that our proposed two-step RGLR analysis delivers notably better results through smaller estimation bias and mean squared error and larger power than the stratified Cox model analysis when there is a treatment-by-stratum interaction, with similar performance when there is no interaction. Additionally, our method controls the type I error rate while the stratified Cox model does not in small samples. We illustrate our method using data from a clinical trial comparing two treatments for colon cancer.

KEYWORDS

minimum risk weights, refined generalized logrank statistic, stratified cox model, treatment-by-stratum interaction, weighted average

1 | INTRODUCTION

In randomized clinical trials with a time-to-event endpoint, it is important to incorporate stratification when the risk of having the event of interest is expected to be influenced by one or more prognostic factors, such as gender, baseline disease severity, specific genetic mutation (eg, HER2 positivity in breast cancer), and so on. Several studies have shown that omitting important covariates from the analysis model can lead to potentially spurious results.[1-5] For example, Schumacher et al[4] showed that the estimated hazard ratio is attenuated if a prognostic factor is omitted, and this result is also confirmed by Bretagnolle and Huber-Carol.[5] A commonly used approach for analyzing stratified trials with time-to-event outcomes is the stratified Cox proportional hazard model,[6] which makes the assumption of proportional hazards within each stratum. It also imposes an additional assumption that the hazard ratio is exactly the same across all strata, which

seems implausible in many practical settings. When there is a treatment by stratum interaction, ie, the hazard ratio differs by stratum, using the conventional stratified Cox model analysis can lead to a biased and/or less efficient result.

To ensure unbiased and efficient results even when there exists a treatment by stratum interaction, Mehrotra et al[7] proposed a two-step approach to allow for different hazard ratios across strata. Their procedure entails fitting a Cox model separately for each stratum and then combining the stratum-specific log hazard ratio estimates to obtain an estimate of the overall log hazard ratio; the latter is defined later in this article and is presumed to be the parameter of interest. They considered two weighting schemes: sample size (SS) weights and minimum risk (MR) weights[8]; both of these are described in the next section. The Mehrotra et al[7] two-step method was developed for large sample applications, as endorsed by Beisel et al[9] based on extensive simulations; however, many randomized clinical trials involve relatively small samples (50-200 patients per treatment group).[10] In the study of Xu et al,[11] we developed a method for improving hazard ratio estimation using a refined generalized logrank (RGLR) statistic for small randomized clinical trials without stratification, and showed that it provides higher efficiency and smaller bias than the Cox proportional hazards model analysis. In this article, we extend the RGLR method to handle stratification and explore its performance in small samples. An additional contribution is the theoretical development of a (remarkably accurate) approximation for the variance of the RGLR-based estimate of a log hazard ratio. Section 2 includes details of the two-step RGLR approach for both the SS and MR weighting schemes. In Section 3, we explore the relative performance of the competing methods, namely, the conventional stratified Cox model analysis and the corresponding two-step Cox model and two-step RGLR analyses, through simulations. We then apply the methods to a real data example from a colon cancer clinical trial in Section 4. Section 5 includes concluding remarks.

2 | METHODS

Suppose there are S strata and within stratum i, where $i = 1, 2, \ldots, S$; we randomize n_{iA} and n_{iB} subjects to treatment A and B, respectively; by design, the ratio n_{iA}/n_{iB} is kept constant across strata. Denote the total SS in stratum i as $n_i = n_{iA} + n_{iB}$ and total SS as $n = \sum_{i=1}^{S} n_i$. Within stratum i, let $t_{i,1} < t_{i,2} < \ldots < t_{i,k_i}$ denote the ordered observed event times for the combined group across treatments. Let $\beta_i = \log[h_{iA}(t)/h_{iB}(t)]$ denote the true time-invariant log hazard ratio in stratum i with $\theta_i = \exp(\beta_i)$ representing the corresponding true hazard ratio and $h_{iA}(t), h_{iB}(t)$ representing the hazard rate in treatment A and B, respectively. If there is no treatment by stratum interaction, ie, if $\beta_i = \beta$ for all i, there is no ambiguity about the definition of the overall log hazard ratio. However, in the presence of an interaction, ie, if $\beta_i \neq \beta_{i*}$ for at least one i and i^*, it is natural to define the target parameter as a population weighted mean of the β_i's, ie, $\bar{\beta} = \sum_{i=1}^{S} f_i \beta_i$, where f_i is the fraction of subjects in the target population that belong to stratum i ($\sum_{i=1}^{S} f_i = 1$). The overall hazard ratio is defined as $\theta_{overall} = \exp(\bar{\beta})$. Some readers may choose to define the true overall hazard ratio as the weighted arithmetic mean of the true stratum-specific hazard ratios. Our preferred choice essentially replaces the arithmetic mean with the geometric mean in such a construct. We do this based on the simple observation that the geometric mean of $\{h_{iA}(t)/h_{iB}(t); i = 1, 2, \ldots, S\}$ always equals the geometric mean of $\{h_{iA}(t); i = 1, 2, \ldots, S\}$ divided by the geometric mean of $\{h_{iB}(t); i = 1, 2, \ldots, S\}$. This appealing feature is lost if the arithmetic mean is used instead. Nevertheless, we stress that the two-step estimation approach described later in this article has the flexibility to accommodate both definitions of the true overall hazard ratio. Moreover, there is no requirement for the true stratum-specific hazard ratios to be equal under either definition. We acknowledge that while this is true for the *definition* of the true overall hazard ratio, its *interpretation* can be challenging if the true stratum-specific hazard ratios are considerably different.

The conventional stratified Cox model analysis assumes no treatment by stratum interaction, and this can (and often does) result in a biased estimate of $\bar{\beta}$, for reasons articulated in Mehrotra et al[7] To allow for a potential treatment by stratum interaction, we propose to use RGLR to estimate the log hazard ratio in each stratum and combine the stratum-specific point estimates using a weighted average to estimate the overall log hazard ratio:

$$\hat{\bar{\beta}} = \sum_{i=1}^{S} \hat{w}_i \hat{\beta}_i. \tag{2.1}$$

Following Mehrotra et al,[7] we consider two weighting schemes: SS and MR. Sample size weighting uses the SS in each stratum relative to the whole sample as the weight, ie, $\hat{w}_i^{SS} = n_i/n$; we assume that $n_i/n \approx f_i$. While SS weighting provides

an unbiased estimator of $\bar{\beta}$, it can suffer from a needlessly large variance. The MR weights proposed by Mehrotra and Railkar[8] in a different context are intended to minimize mean squared error (MSE); for our stratified time-to-event setting, the weights are calculated as follows:

$$\hat{w}_i^{MR} = \frac{a_i}{\sum_{i=1}^{S} \hat{V}_i^{-1}} - \frac{b_i \hat{V}_i^{-1}}{\sum_{i=1}^{S} \hat{V}_i^{-1} + \sum_{i=1}^{S} b_i \hat{\beta}_i \hat{V}_i^{-1}} \cdot \frac{\sum_{i=1}^{S} \hat{\beta}_i a_i}{\sum_{i=1}^{S} \hat{V}_i^{-1}}, \tag{2.2}$$

where $b_i = \hat{\beta}_i \sum_{i=1}^{S} \hat{V}_i^{-1} - \sum_{i=1}^{S} \hat{\beta}_i \hat{V}_i^{-1}$, $a_i = \hat{V}_i^{-1}(1 + b_i \sum_{i=1}^{S} \hat{\beta}_i n_i / n)$, and $\hat{V}_i$ is the estimated variance for $\hat{\beta}_i$. Of note, when the log hazard ratios are approximately the same across all strata, it is easily seen that $\hat{w}_i^{MR} \approx \hat{V}_i^{-1}$, which is intuitively appealing because the latter weights are optimal (and very similar to the weights used implicitly by the stratified Cox model) when there is no treatment by stratum interaction; see Mehrotra et al.[7]

To implement Equation 2.2, we need to derive the variance of the stratum-specific RGLR estimate of the log hazard ratio. Details of the RGLR definition and derivation can be found in our previous work.[11] We will provide the basic set up of RGLR here. Consider the scenario of single stratum, ie, $S = 1$. Let T denote the random variable for the event time. Then, by definition, we have $P(t_{i,j-1} < T \leq t_{i,j} | T > t_{i,j-1}) = 1 - \exp(-p_{i,j})$, where $p_{i,j} = \int_{t_{i,j-1}}^{t_{i,j}} h_{iB}(x)\mathrm{d}x$, and $h_{iB}(t)$ is the hazard function for group B for stratum i. Let random variables $D_{i,jA}$ and $D_{i,jB}$ denote the number of events at time $t_{i,j}$ in group A and B, respectively, and let $D_{i,j} = D_{i,jA} + D_{i,jB}$. Let random variables $R_{i,jA}$ and $R_{i,jB}$ denote the number of subjects at risk at time $t_{i,j}$ in group A and B, respectively, and $r_{i,jA}, r_{i,jB}$ denote the corresponding observed number of subjects. Under the assumption of no tied event times, given $d_{i,j}, r_{i,jA}, r_{i,jB}, p_{i,j}, \beta_i$, $D_{i,jA}$ follows a noncentral hypergeometric distribution with probability $1 - \exp(-e^{\beta_i} p_{i,j})$ and $R_{i,jA}$ number of trials. The conditional mean and variance of $D_{i,jA}$ can then be derived as follows:

$$E_{i,jA} = \frac{r_{i,jA}(1 - e^{-p_{i,j}e^{\beta_i}})e^{-p_{i,j}}}{r_{i,jA}(1 - e^{-p_{i,j}e^{\beta_i}})e^{-p_{i,j}} + r_{i,jB}(1 - e^{-p_{i,j}})e^{-p_{i,j}e^{\beta_i}}} \tag{2.3}$$

$$V_{i,jA} = \frac{r_{i,jA}(1 - e^{-p_{i,j}e^{\beta_i}})e^{-p_{i,j}} r_{i,jB}(1 - e^{-p_{i,j}})e^{-p_{i,j}e^{\beta_i}}}{[r_{i,jA}(1 - e^{-p_{i,j}e^{\beta_i}})e^{-p_{i,j}} + r_{i,jB}(1 - e^{-p_{i,j}})e^{-p_{i,j}e^{\beta_i}}]^2}. \tag{2.4}$$

We showed in Xu et al[11] that the nuisance parameter can be estimated using an unconditional approach, where

$$\tilde{p}_{i,j} = \begin{cases} \log\left(\frac{\theta R_{i,jA} + R_{i,jB}}{\theta R_{i,jA} + R_{i,jB} - 1}\right), \text{ when } d_{i,jA} = 0, d_{i,jB} = 1 \\ \log\left(\frac{\theta R_{i,jA} + R_{i,jB}}{\theta R_{i,jA} + R_{i,jB} - \theta}\right), \text{ when } d_{i,jA} = 1, d_{i,jB} = 0. \end{cases} \tag{2.5}$$

Let $\tilde{\mathbf{p}}_\mathbf{i}$ denote the estimated nuisance parameter vector and define $S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i}) = \sum_{j=1}^{k}(d_{i,jA} - E_{i,jA})$ and $I_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i}) = \sum_{j=1}^{k} V_{i,jA}$. Note that $E[S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})] = 0$, and therefore, we can estimate the log hazard ratio β_i by solving the moment equation $S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i}) = 0$. Denote the moment log hazard ratio estimator by $\hat{\beta}_i^{RGLR}$. Then, by the first order Taylor expansion, we have

$$\hat{\beta}_i^{RGLR} - \beta_i \cong -\frac{S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\frac{\partial S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\partial \beta_i}}, \tag{2.6}$$

where

$$-\frac{\partial S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\partial \beta_i} = \sum_{j=1}^{k} \frac{p_{i,j} e^{\beta_i} e^{-p_{i,j}e^{-\beta_i}}}{1 - e^{-p_{i,j}e^{\beta_i}}} V_{i,jA}. \tag{2.7}$$

As SS gets large, $p_{i,j} \to 0$, and by L'Hôpital's rule, we have that $\lim_{p_{i,j} \to 0} \frac{p_{i,j} e^{\beta_i} e^{-p_{i,j}e^{-\beta_i}}}{1 - e^{-p_{i,j}e^{\beta_i}}} = 1$. Thus,

$$-\frac{\partial S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\partial \beta_i} \cong \sum_{j=1}^{k} V_{i,jA} = I_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i}). \tag{2.8}$$

Combining Equations 2.6 and 2.8 gives us

$$\sqrt{I_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}(\hat{\beta}_i^{RGLR} - \beta_i) \cong \frac{S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\sqrt{I_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}}. \tag{2.9}$$

Since $S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})$ converges to the Cox partial likelihood score function as $n \to \infty$, and $\tilde{\mathbf{p}}_\mathbf{i}$ goes to 0, an argument similar to that in Andersen and Gill[12] can be applied to show asymptotic normality of the RGLR estimator for β_i. Specifically, by the Martingale Central Limit Theorem,

$$\frac{S_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}{\sqrt{I_{i,k}(\beta_i, \tilde{\mathbf{p}}_\mathbf{i})}} \xrightarrow{D} N(0,1). \tag{2.10}$$

Therefore, denoting $\hat{V}_i(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}_\mathbf{i}) = I_{i,k}^{-1}(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}_\mathbf{i})$ and combining Equations 2.9 and 2.10, we have

$$\frac{\hat{\beta}_i^{RGLR} - \beta_i}{\sqrt{\hat{V}_i(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}_\mathbf{i})}} \xrightarrow{D} N(0,1), \tag{2.11}$$

where

$$\hat{V}_i(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}_\mathbf{i}) \approx \left(\sum_{j=1}^{k} \frac{r_{i,jA}\left(1 - e^{-\tilde{p}_{i,j}e^{\hat{\beta}_i^{RGLR}}}\right) e^{-\tilde{p}_{i,j}} r_{i,jB}(1 - e^{-\tilde{p}_{i,j}}) e^{-\tilde{p}_{i,j}e^{\hat{\beta}_i^{RGLR}}}}{\left[r_{i,jA}\left(1 - e^{-\tilde{p}_{i,j}e^{\hat{\beta}_i^{RGLR}}}\right) e^{-\tilde{p}_{i,j}} + r_{i,jB}(1 - e^{-\tilde{p}_{i,j}}) e^{-\tilde{p}_{i,j}e^{\hat{\beta}_i^{RGLR}}} \right]^2} \right)^{-1}. \tag{2.12}$$

With the approximate variance formula now established for the RGLR estimator in each stratum, we can now calculate the MR weights using Equation 2.2. For both weighting schemes, we do hypothesis testing ($H_0 : \bar{\beta} = 0$ vs $H_1 : \bar{\beta} \neq 0$) using $\hat{\bar{\beta}}^{RGLR}$, the weighted average of the S independent stratum-specific estimates $\hat{\beta}_i^{RGLR}$, as shown in Equation 2.1. The approximate variance of $\hat{\bar{\beta}}^{RGLR}$ is calculated as follows:

$$\hat{V}\left(\hat{\bar{\beta}}^{RGLR}\right) = \sum_{i=1}^{S} \hat{w}_i^2 \hat{V}_i(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}_\mathbf{i}). \tag{2.13}$$

Confidence interval (CI) calculations can be done using Wald tests implied by Equation 2.11. A numerical study demonstrating the impressive accuracy of the variance Formulas 2.12 and 2.13 is provided in the Supporting Information. It is important to note that Equation 2.13 ignores the variability in the estimated weights. As such, strictly speaking, the corresponding inference should be treated as being approximate. Fortunately, as seen later in our simulation results, the approximation is good enough to deliver excellent type 1 error rate and confidence interval coverage properties across a range of realistic scenarios. Researchers who prefer a closer-to-exact inference option can replace use of the approximate variance in Equation 2.13 with a corresponding variance obtained from a standard nonparametric bootstrap (details omitted).

3 | SIMULATIONS

3.1 | Simulation setup

We performed a simulation study to examine the bias, relative efficiency and nominal 95% CI coverage probability of the two-step RGLR using SS weights and MR weights and compared the performance of our proposed methods to the conventional stratified Cox proportional hazards analysis and the two-step method of Mehrotra et al[7] in which stratum-specific Cox model estimates are combined using SS weights or MR weights.

We considered the case of two strata and four strata in the simulation study. Usually, in the presence of stratification, only the total number of subjects per group and randomization ratio (= 1 here) is fixed by design. Therefore, we used a similar simulation setup as Mehrotra and Railkar[8] and treated the number of subjects in each stratum as a random variable. Specifically, n pairs of subjects were first assigned to stratum i with probability f_i ($\sum f_i = 1$), where $i = 1, 2$ for two strata and $i = 1, 2, 3, 4$ for four strata, and then, within each pair, one subject was randomly assigned to treatment A and the other to treatment B with equal probability. Thereafter, for subject j in stratum i and randomized to treatment q (q =A or B), we generated an entry time e_{ijq} from a uniform distribution (0, T). For two strata, survival times s_{ijq} for subject j under treatment A and treatment B in stratum i were generated from Weibull (scale= $\lambda_i/\sqrt{\theta_i}$, shape = 2) and Weibull (scale = λ_i, shape = 2) respectively, where $\lambda_1 = 0.6, \lambda_2 = 1.2$. Note that the hazard function for Weibull (scale = λ, shape = γ) is $\gamma x^{\gamma-1}/\lambda^\gamma$, so the hazard ratio of treatment A relative to B in stratum i is θ_i. The follow-up time for a subject j randomized to treatment q in stratum i was $t_{ijq} = \min(s_{ijq}, T - e_{ijq})$.

TABLE 1 True log hazard ratio in each stratum and overall under the null and alternative hypotheses

Two Strata				
Scenario 1: Equal stratum sizes				
Stratum	**Relative frequency**	**Null (no interaction)**	**Alternative 1 (no interaction)**	**Alternative 2 (interaction)**
1	0.5	0	-0.7	-0.2
2	0.5	0	-0.7	-1.2
Overall		0	-0.7	-0.7
Scenario 2: Unequal stratum sizes				
Stratum	**Relative frequency**	**Null (no interaction)**	**Alternative 1 (no interaction)**	**Alternative 2 (interaction)**
1	0.7	0	-0.7	-0.4
2	0.3	0	-0.7	-1.4
Overall		0	-0.7	-0.7
Four Strata				
Scenario 1: Equal stratum sizes				
Stratum	**Relative frequency**	**Null (no interaction)**	**Alternative 1 (no interaction)**	**Alternative 2 (interaction)**
1	0.25	0	-0.7	-0.3
2	0.25	0	-0.7	-0.4
3	0.25	0	-0.7	-0.8
4	0.25	0	-0.7	-1.3
Overall		0	-0.7	-0.7
Scenario 2: Unequal stratum sizes				
Stratum	**Relative frequency**	**Null (no interaction)**	**Alternative 1 (no interaction)**	**Alternative 2 (interaction)**
1	0.15	0	-0.7	-0.3
2	0.35	0	-0.7	-0.4
3	0.35	0	-0.7	-0.8
4	0.15	0	-0.7	-1.65
Overall		0	-0.7	-0.7

Under all the alternative hypotheses for both two strata and four strata, the overall log hazard ratio $\bar{\beta}$ is fixed at -0.7.

For four strata, we used the same procedure for generating number of subjects per stratum, entry time, and survival time as described above for the two strata simulations. Survival time s_{ijq} for subject j in stratum i under treatment A and B was generated from Weibull (scale $= \lambda_i/\sqrt{\theta_i}$, shape $= 2$) and Weibull (scale $= \lambda_i$, shape $= 2$), respectively, where now with $\lambda_1 = 0.6, \lambda_2 = 0.8, \lambda_3 = 1$, and $\lambda_4 = 1.2$.

We varied the stratum-specific relative frequency and true log hazard ratio, along with total SS and overall percentage censoring. Both equal (scenario 1) and unequal (scenario 2) stratum sizes were considered. For two strata, we set $f_1 = f_2 = 0.5$ and $f_1 = 0.7, f_2 = 0.3$ for scenarios 1 and 2, respectively. For four strata, we set $f_1 = f_2 = f_3 = f_4 = 0.25$ and $f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$ for scenarios 1 and 2, respectively. Under the null hypothesis, stratum-specific and overall log hazard ratio were 0 in all cases. Under the alternative hypothesis, we considered two settings: the same log hazard ratio across strata (Alt 1) and different log hazard ratios across strata (Alt 2). The stratum-specific log hazard ratios in each scenario are summarized in Table 1; of note, the overall log hazard ratio ($\bar{\beta}$) was fixed at -0.7 in every case, which corresponds to an overall hazard ratio of $\exp(-0.7) = 0.5$. Subjects per treatment group was varied as 50, 100 for two strata, and 100, 200 for four strata. Two percentage censoring values were considered, both controlled by selecting a specific T conditional on the fixed β_i's for the given scenario: 25% and 50%. 5000 replications were generated. Hypothesis testing was done at the $\alpha = 0.05$ level. Results for bias (under the null hypothesis), percent bias (under the alternative hypothesis), type I error rate, power, relative efficiency, and coverage probability for the 95% CI for two and four strata were obtained. Here, relative efficiency refers to 100 times the ratio of the MSE) for the estimator of $\bar{\beta}$ using the stratified Cox model relative to that using the given alternative method of estimation. Thus, relative efficiency greater than 100% represents an improvement over the stratified Cox model.

3.2 | Simulation results

Table 2 shows the results for the two strata case under the null hypothesis and the two alternative hypotheses for both equal (scenario 1) and unequal (scenario 2) relative frequency in each stratum. In scenario 1, under the null hypothesis, all methods were associated with negligible bias. Our proposed two-step RGLR provided similar efficiency relative to

TABLE 2 Bias (% bias), percent relative efficiency defined as 100 times the ratio of mean squared error for the stratified Cox model relative to competing method and coverage probability for 95% CI for overall log hazard ratio $\bar{\beta}$ for two strata based on 5000 simulations*

Scenario 1: Equal Stratum Sizes ($f_1 = f_2 = 0.5$)										
			Null			**Alt 1 (no interaction)**			**Alt 2 (interaction)**	
Censoring	***N*/trt**	**Method**	**Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**
25%	50	Stratified Cox	-0.001	100	(94.2)	1.8	100	94.6	-12.7	100	(92.8)
		Two-step Cox (SS wts)	-0.001	95	(93.9)	3.7	94	(94.0)	4.2	93	(94.1)
		Two-step RGLR (SS wts)	-0.001	102	94.7	0.1	101	95.0	0.5	100	94.8
		Two-step Cox (MR wts)	-0.001	97	(93.9)	2.8	97	(94.3)	0.7	97	(94.3)
		Two-step RGLR (MR wts)	-0.001	105	94.9	-0.8	104	95.1	-3.0	103	94.7
	100	Stratified Cox	-0.001	100	95.0	0.9	100	94.8	-13.5	100	(90.1)
		Two-step Cox (SS wts)	-0.002	97	94.8	1.8	97	94.5	2.5	110	(94.0)
		Two-step RGLR (SS wts)	-0.002	102	95.3	-0.2	100	95.0	0.5	115	94.4
		Two-step Cox (MR wts)	-0.001	99	94.8	1.54	98	94.5	0.5	113	(94.1)
		Two-step RGLR (MR wts)	-0.001	103	95.3	-0.6	102	95.0	-1.5	117	94.4
50%	50	Stratified Cox	0.000	100	94.4	2.0	100	94.7	-27.1	100	(89.8)
		Two-step Cox (SS wts)	0.001	86	94.7	4.3	85	95.1	4.8	97	95.7
		Two-step RGLR (SS wts)	0.001	93	95.6	0.4	93	96.0	1.0	106	96.3
		Two-step Cox (MR wts)	0.000	94	94.4	2.8	93	94.8	-4.3	109	94.5
		Two-step RGLR (MR wts)	0.000	102	95.5	-1.0	102	95.3	-8.3	116	94.8
	100	Stratified Cox	-0.001	100	95.0	1.1	100	94.9	-28.3	100	(82.7)
		Two-step Cox (SS wts)	-0.003	89	94.8	2.4	89	94.6	2.9	135	94.9
		Two-step RGLR (SS wts)	-0.003	93	95.3	0.3	93	95.1	0.8	142	95.2
		Two-step Cox (MR wts)	-0.002	95	94.7	1.7	95	94.6	-3.0	141	(93.5)
		Two-step RGLR (MR wts)	-0.002	99	95.1	-0.4	99	94.9	-5.2	145	(93.6)
Scenario 2: Unequal Stratum Sizes ($f_1 = 0.7, f_2 = 0.3$)										
			Null			**Alt 1 (no interaction)**			**Alt 2 (interaction)**	
Censoring	***N*/trt**	**Method**	**Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**
25%	50	Stratified Cox	0.002	100	(94.2)	1.6	100	94.6	-12.5	100	(93.4)
		Two-step Cox (SS wts)	0.003	93	(94.2)	3.5	92	94.3	4.6	91	(94.2)
		Two-step RGLR (SS wts)	0.002	101	94.9	0.0	99	95.0	0.0	100	94.9
		Two-step Cox (MR wts)	0.002	96	(94.3)	2.6	96	94.5	0.2	98	(94.2)
		Two-step RGLR (MR wts)	0.002	104	94.7	-0.9	103	95.0	-4.4	105	94.4
	100	Stratified Cox	0.003	100	95.3	0.6	100	95.5	-12.1	100	(91.3)
		Two-step Cox (SS wts)	0.002	96	95.0	1.5	96	95.3	2.5	106	94.4
		Two-step RGLR (SS wts)	0.002	101	95.6	-0.5	100	95.7	0.0	112	94.8
		Two-step Cox (MR wts)	0.002	98	95.0	1.1	98	95.4	0.1	111	94.4
		Two-step RGLR (MR wts)	0.002	103	95.5	-0.9	102	95.5	-2.4	115	94.7
50%	50	Stratified Cox	-0.001	100	95.0	1.5	100	95.0	-21.7	100	(90.0)
		Two-step Cox (SS wts)	-0.003	87	95.2	3.2	89	95.0	-0.3	104	95.9
		Two-step RGLR (SS wts)	-0.003	95	96.1	-0.5	97	95.8	-4.7	113	96.2
		Two-step Cox (MR wts)	-0.002	95	95.0	2.0	96	94.7	-8.4	112	(94.3)
		Two-step RGLR (MR wts)	-0.002	103	95.7	-1.6	104	95.5	-12.8	116	94.4
	100	Stratified Cox	0.004	100	95.2	0.4	100	95.1	-21.0	100	(87.9)
		Two-step Cox (SS wts)	0.004	88	95.0	1.3	89	94.9	4.1	106	95.5
		Two-step RGLR (SS wts)	0.004	92	95.7	-0.7	93	95.3	1.4	113	95.9
		Two-step Cox (MR wts)	0.004	94	94.9	0.8	95	94.8	-2.3	116	(94.3)
		Two-step RGLR (MR wts)	0.003	99	95.3	-1.2	99	95.3	-5.1	121	94.4

Abbreviations: Alt, alternative; Cov, coverage; MR wts, minimum risk weights; RGLR, refined generalized logrank statistic; SS wts, sample size weights; Trt, treatment group. *Bias is reported under the null hypothesis and percentage bias is reported under the alternative hypothesis. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. Each two-step method uses a weighted average of stratum-specific log hazard ratio estimates.

the stratified Cox model and higher efficiency than the Mehrotra et al[7] two-step Cox model method under both weighting schemes. Our proposed method also controlled the type I error rate under 5% across all simulated scenarios, while both the stratified Cox model and the two-step Cox model method had inflated type I error for 50 subjects per treatment group and 25% censoring. Under the alternative hypothesis with no stratum by treatment interaction (Alt 1), the stratified Cox is expected to have the best performance, and the two-step RGLR provided very similar efficiency relative to the stratified Cox model. The two-step RGLR also delivered a percentage bias less than 2% and maintained adequate

coverage probability for the 95% CI, while the stratified Cox model failed to do so under equal stratum SS with 50 subjects per treatment and 25% censoring. When there was interaction between treatment and stratum (Alt 2), the proposed two-step RGLR provided notably better efficiency and smaller bias than all the other competing methods. Both the stratified and two-step Cox model methods had issues with maintaining adequate 95% CI coverage probability in several simulated scenarios, but the two-step RGLR with SS weights maintained adequate coverage probability throughout all simulated settings. The two-step RGLR with MR weights also performed well but it failed to maintain adequate coverage probability in the scenario with 100 subjects per treatment and 50% censoring. With 100 subjects per treatment and 50% censoring, the two-step RGLR with SS weights delivered 42% higher efficiency than the stratified Cox model, with

TABLE 3 Bias (% bias), percent relative efficiency defined as 100 times the ratio of mean squared error for the stratified Cox model relative to competing method and coverage probability for 95% CI for overall log hazard ratio $\bar{\beta}$ for 4 strata based on 5000 simulations*

Scenario 1: Equal Stratum Sizes ($f_1 = f_2 = f_3 = f_4 = 0.25$)											
			Null			**Alt 1 (no interaction)**			**Alt 2 (interaction)**		
Censoring	***N*/trt**	**Method**	**Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**
25%	100	Stratified Cox	0.000	100	94.8	0.7	100	95.2	-8.5	100	(93.1)
		Two-step Cox (SS wts)	0.001	93	(94.0)	3.3	91	94.6	3.6	92	(94.0)
		Two-step RGLR (SS wts)	0.001	101	94.8	-0.3	99	95.3	-0.1	100	94.8
		Two-step Cox (MR wts)	0.001	95	(94.1)	2.6	94	94.9	1.8	95	(94.1)
		Two-step RGLR (MR wts)	0.001	103	94.9	-1.0	101	95.4	-1.9	102	94.6
	200	Stratified Cox	0.001	100	94.8	0.4	100	95.2	-9.2	100	(91.6)
		Two-step Cox (SS wts)	0.001	97	94.7	1.6	96	94.9	1.6	114	(94.2)
		Two-step RGLR (SS wts)	0.001	102	95.2	-0.4	100	95.4	-0.4	119	94.7
		Two-step Cox (MR wts)	0.001	98	94.7	1.3	97	95.0	0.6	116	(94.1)
		Two-step RGLR (MR wts)	0.001	103	95.1	-0.7	101	95.3	-1.5	119	94.4
50%	100	Stratified Cox	-0.001	100	94.9	1.1	100	94.7	-16.0	100	(90.7)
		Two-step Cox (SS wts)	-0.001	87	94.5	4.1	85	94.5	4.3	94	94.9
		Two-step RGLR (SS wts)	-0.001	94	95.4	0.2	93	95.3	0.4	104	95.7
		Two-step Cox (MR wts)	-0.001	92	(94.3)	2.9	90	(94.2)	0.2	102	(93.9)
		Two-step RGLR (MR wts)	-0.001	99	95.3	-1.0	99	94.9	-3.8	110	94.6
	200	Stratified Cox	0.000	100	94.9	0.4	100	95.2	-16.9	100	(86.4)
		Two-step Cox (SS wts)	-0.001	94	94.9	1.9	91	94.8	2.2	129	95.0
		Two-step RGLR (SS wts)	-0.001	98	95.4	-0.3	96	95.4	0.1	136	95.3
		Two-step Cox (MR wts)	-0.001	96	94.7	1.3	95	94.6	-0.5	134	(94.3)
Scenario 2: Unequal Stratum Sizes ($f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$)											
			Null			**Alt 1 (no interaction)**			**Alt 2 (interaction)**		
Censoring	***N*/trt**	**Method**	**Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**	**%Bias**	**%RE**	**Cov**
25%	100	Stratified Cox	0.000	100	95.1	0.8	100	95.6	-9.8	100	(92.4)
		Two-step Cox (SS wts)	0.000	93	94.2	3.5	91	94.9	3.4	95	(94.3)
		Two-step RGLR (SS wts)	0.000	101	95.0	-0.1	99	95.6	-0.6	104	94.9
		Two-step Cox (MR wts)	0.000	95	94.4	2.6	94	95.2	1.2	100	(94.3)
		Two-step RGLR (MR wts)	0.000	103	95.1	-0.9	101	95.5	-2.8	107	94.6
	200	Stratified Cox	-0.000	100	95.5	0.5	100	95.2	-9.8	100	(91.1)
		Two-step Cox (SS wts)	-0.001	97	95.2	1.7	96	94.9	2.0	113	94.9
		Two-step RGLR (SS wts)	-0.000	101	95.6	-0.2	100	95.2	-0.2	119	95.3
		Two-step Cox (MR wts)	-0.000	98	95.1	1.4	97	94.8	0.9	116	94.8
		Two-step RGLR (MR wts)	-0.000	102	95.8	-0.6	101	95.4	-1.4	120	95.1
50%	100	Stratified Cox	0.002	100	95.0	0.5	100	95.3	-16.8	100	(90.9)
		Two-step Cox (SS wts)	0.002	91	94.8	3.4	89	95.2	-0.3	112	95.4
		Two-step RGLR (SS wts)	0.002	98	95.7	-0.4	97	95.9	-4.2	119	95.9
		Two-step Cox (MR wts)	0.002	95	94.6	2.2	94	94.9	-4.1	116	94.7
		Two-step RGLR (MR wts)	0.002	102	95.4	-1.6	102	95.6	-8.0	119	94.7
	200	Stratified Cox	0.001	100	95.0	0.2	100	95.4	-16.5	100	(86.8)
		Two-step Cox (SS wts)	0.002	93	94.9	1.7	93	95.2	1.9	128	95.4
		Two-step RGLR (SS wts)	0.002	97	95.3	-0.4	97	95.4	-0.4	136	95.5
		Two-step Cox (MR wts)	0.002	96	94.6	1.2	96	95.0	-0.9	134	94.4
		Two-step RGLR (MR wts)	0.002	100	95.1	-0.9	100	95.2	-3.2	138	94.7

Abbreviations: Alt, alternative; Cov, coverage; MR wts, minimum risk weights; RGLR, refined generalized logrank statistic; SS wts, sample size weights; Trt, treatment group. *Bias is reported under the null hypothesis and percentage bias is reported under the alternative hypothesis. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. Each two-step method uses a weighted average of stratum-specific log hazard ratio estimates.

a percentage bias of 0.8%, comparing to -28.3% bias from the stratified Cox model. The performance of the methods for unequal relative frequency in each stratum was similar to that for equal relative frequency described above.

Table 3 shows the results for the four strata case. Under both equal and unequal relative stratum frequency, our two-step RGLR provided the smallest bias and higher relative efficiency compared to the stratified Cox model. When there was a treatment by stratum interaction, the stratified Cox model had a bias as large as -16.9%, while the two-step RGLR controlled the bias under 8%. In terms of type I error, the stratified and two-step Cox model methods had inflated type I error issues with smaller SSs (100 subjects per treatment group with 25% and 50% censoring under scenario 1), while our two-step RGLR did not. In terms of coverage probability, the two-step RGLR maintained adequate coverage probability for 95% CI throughout all scenarios, while the stratified Cox model failed to do so under several scenarios.

We also examined power among the methods. Table 4 shows the results for 100 subjects per treatment with 50% censoring for two strata and four strata cases; results under other simulated scenarios (not shown) did not provide additional insights and are hence not shown. When there was no interaction between treatment and stratum, our two-step RGLR provided similar power as the stratified Cox model. When there was interaction, using two-step RGLR delivered a power increase of at least 5 percentage points relative to the stratified Cox model. While the two-step Cox model method seemed to have slightly better power than the two-step RGLR, the former also sometimes had inflated type I error rate while our two-step RGLR did not.

4 | APPLICATION

We apply the stratified Cox model, the Mehrotra et al[7] two-step Cox model method and our proposed two-step RGLR method, with both two-step methods using SS and MR weights, to a clinical trial involving resected colon cancer.[13] The data set included 154 patients with stage C colon cancer who were randomized to receive placebo or levamisole combined with fluorouracil therapy, with 77 patients in each group. The outcome of interest was overall survival. Patients were stratified by the number of lymph nodes involved ($\leq$ 4 vs >4). Table 5 summarizes the results from applying all the methods. The stratified Cox model provided an estimated overall hazard ratio (therapy:placebo) of $\exp(-0.64) = 0.53$ (95% CI, 0.31-0.90), with a P value of 0.021. On the other hand, the two-step Cox and two-step RGLR, for both SS and MR weights, provided a non-significant P value (>0.05). The estimated hazard ratio in stratum 1 using Cox and RGLR was $\exp(-0.27) = 0.76$ and $\exp(-0.26) = 0.77$, respectively, with corresponding estimates of the hazard ratio in stratum 2 being $\exp(-1.16) = 0.31$ and $\exp(-1.14) = 0.32$, respectively. The MR weight for both two-step RGLR and two-step Cox is 0.69 and 0.31 for strata 1 and 2, respectively. The SS weight is 0.73 and 0.27 for strata 1 and 2, respectively. The Kaplan-Meier curves by stratum in Figure 1 appear to support a differential treatment effect across the two strata; ie, they suggest evidence of a treatment by stratum interaction, thereby casting doubt on the stratified Cox model analysis, which

TABLE 4 Power comparisons among the competing methods based on 100 subjects per treatment group and 50% censoring with 5000 simulations for two strata (top panel) and four strata (bottom panel)

	Two Strata			
	Scenario 1: $f_1 = f_2 = 0.5$		**Scenario 2:** $f_1 = 0.7, f_2 = 0.3$	
Method	**Alt 1 (no interaction)**	**Alt 2 (interaction)**	**Alt 1 (no interaction)**	**Alt 2 (interaction)**
Stratified Cox	92.5	(66.8)	96.2	(80.5)
Two-step Cox (SS wts)	90.4	86.2	94.9	90.7
Two-step RGLR (SS wts)	89.8	85.0	94.4	89.5
Two-step Cox (MR wts)	91.8	(84.2)	95.6	(89.9)
Two-step RGLR (MR wts)	91.3	(83.1)	95.1	88.9
	Four Strata			
	Scenario 1: $f_1 = f_2 = f_3 = f_4 = 0.25$		**Scenario 2:** $f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$	
Method	**Alt 1 (no interaction)**	**Alt 2 (interaction)**	**Alt 1 (no interaction)**	**Alt 2 (interaction)**
Stratified Cox	91.2	78.8	90.8	(80.6)
Two-step Cox (SS wts)	89.8	86.9	89.9	87.3
Two-step RGLR (SS wts)	88.5	85.4	88.4	85.6
Two-step Cox (MR wts)	(90.9)	(87.2)	90.8	87.4
Two-step RGLR (MR wts)	89.7	85.5	89.3	85.6

Abbreviations: Alt, alternative; MR wts, minimum risk weights; RGLR, refined generalized logrank; SS wts, sample size weigthts. Square brackets indicate the case where the coverage probability is more than $Z_{0.975}$ standard errors below 95%.

TABLE 5 Log hazard ratio estimates for the colon cancer data example in Lin et al.[13]

	N(%)	Stratified Cox	Two-Step Cox (SS wts)	Two-Step RGLR (SS wts)	Two-Step Cox (MR wts)	Two-Step RGLR (MR wts)
Stratum 1 $\hat{\beta}_1$	112 (73%)	-0.64*	-0.27	-0.26	-0.27	-0.26
Stratum 2 $\hat{\beta}_2$	42 (27%)	-0.64*	-1.16	-1.14	-1.16	-1.14
$\hat{\bar{\beta}}$		-0.64*	-0.51	-0.50	-0.54	-0.53
95% CI		(-1.18 to -0.10)	(-1.08 to 0.05)	(-1.07 to 0.06)	(-1.10 to 0.02)	(-1.09 to 0.02)
P-value		0.021	0.075	0.080	0.057	0.060

Abbreviations: CI, confidence interval; MR wts, minimum risk weights; RGLR, refined generalized logrank statistic; SS wts, sample size weights. *The stratified Cox model assumes $\beta_1 = \beta_2$; these are the implied stratum-specific estimates based on the overall estimate.

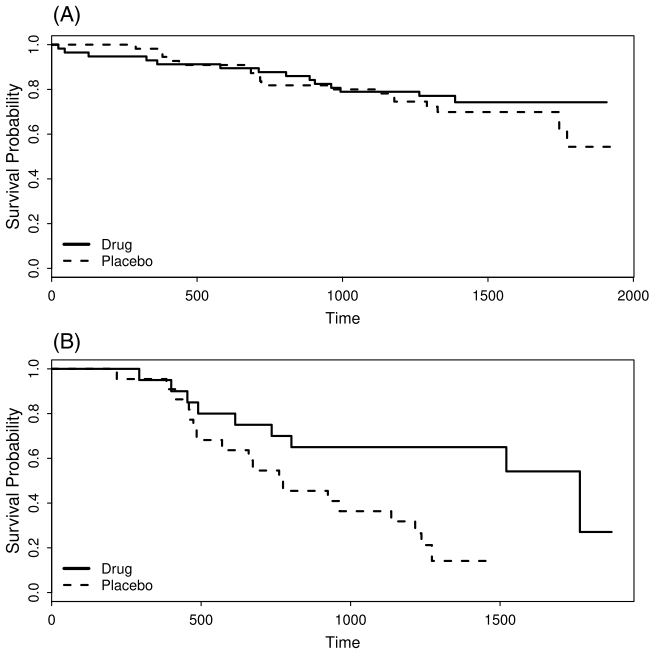

FIGURE 1 Kaplan-Meier survival curves by treatment group; A, is for stratum 1 B, is stratum 2

assumes no treatment by stratum interaction. The overall hazard ratio from the two-step RGLR with SS and MR weights was estimated to be $\exp(-0.50) = 0.61$ (95% CI, 0.34-1.06) and $\exp(-0.53) = 0.59$ (95% CI, 0.83-1.02), respectively. These estimates are deemed more reliable than those based on the stratified Cox model.

5 | CONCLUSIONS

The stratified Cox model is often used to analyze stratified randomized clinical trials with time-to-event data. However, the assumption of equal hazard ratios across strata may not be true in real applications. Therefore, it is important to develop methods to handle a treatment by stratum interaction, especially in relatively small stratified trials with low power to detect a treatment by stratum interaction. In this work, we proposed a two-step approach in which we estimate stratum-specific log hazard ratios using the RGLR approach and combine them across strata using SS or MR weights. Through simulation studies, we have shown that the two-step RGLR provides notably smaller bias and smaller MSE than the conventional stratified Cox model when there is a treatment-by-stratum interaction, with similar performance when there is no interaction. The stratified Cox model tends to inflate the type I error in small samples, while the two-step RGLR does not. The stratified Cox model also has issues with CI under-coverage in small samples, while the two-step RGLR with SS weights does not and with MR weights generally does not. The two-step RGLR method also delivers much higher power than the stratified Cox model when the hazard ratio differs across strata while suffering no material power loss in other cases. Finally, the proposed method has similar or better performance than the two-step method of Mehrotra et al[7] in terms of bias and MSE; this to be be expected because within each stratum, the RGLR estimator outperforms the Cox model estimator in small to moderate SSs, notably so in small samples.[11]

The two-step RGLR removes the restrictive assumption of equal hazard ratios across strata in the stratified Cox model analysis and outperforms the stratified Cox model when there is an interaction between treatment and stratum. More importantly, the two-step RGLR also provides an estimated stratum-specific hazard ratio, while the stratified Cox model only provides an estimated overall hazard ratio. As shown in the colon cancer example, when the hazard ratio is different across strata, using the two-step RGLR can provide additional insight into the difference across strata, while the stratified Cox model does not.

ORCID

Rengyi Xu https://orcid.org/0000-0003-4135-668X

Devan V. Mehrotra https://orcid.org/0000-0002-0316-7362

REFERENCES

1. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Stat Med*. 2002;21(19):2917-2930.
2. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Stat Med*. 2002;21(19):2899-2908.
3. Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika*. 1986;73(2):363-369.
4. Schumacher M, Olschewski M, Schmoor C. The impact of heterogeneity on the comparison of survival times. *Stat Med*. 1987;6(7):773-784.
5. Bretagnolle J, Huber-Carol C. Effects of omitting covariates in Cox's model for survival data. *Scand J Stat*. 1988;15(2):125-138.
6. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc*. 1972;34:187-220.
7. Mehrotra DV, Su S-C, Li X. An efficient alternative to the stratified cox model analysis. *Stat Med*. 2012;31(17):1849-1856.
8. Mehrotra DV, Railkar R. Minimum risk weights for comparing treatments in stratified binomial trials. *Stat Med*. 2000;19(6):811-825.
9. Beisel C, Benner A, Kunz C, Kopp-Scheneider A. Heterogeneous treatment effects in stratified clinical trials with time-to-event outcomes. *Biom J*. 2017;59(3):511-530.
10. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester, West Sussex, England: John Wiley & Sons; 1983.
11. Xu R, Shaw PA, Mehrotra DV. Hazard ratio estimation in small samples. *Stat Biopharmaceutical Res*. 2018;10(2):139-149.
12. Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Ann Stat*. 1982;10(4):1100-1120.
13. Lin D-Y, Dai L, Cheng G, Sailer MO. On confidence intervals for the hazard ratio in randomized clinical trials. *Biometrics*. 2016;72(4):1098-1102.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Xu R, Mehrotra DV, Shaw PA. Hazard ratio inference in stratified clinical trials with time-to-event endpoints and limited sample size. *Pharmaceutical Statistics*. 2019;1–11. https://doi.org/10.1002/pst.1928