BMC Medical Research Methodology

RESEARCH

Open Access

Comparison between asymptotic and re-randomisation tests under non-proportional hazards in a randomised controlled trial using the minimisation method

Ryusei Kimura[1,2*], Shogo Nomura[3], Kengo Nagashima[1] and Yasunori Sato[1,4]

Abstract

Background Pocock-Simon's minimisation method has been widely used to balance treatment assignments across prognostic factors in randomised controlled trials (RCTs). Previous studies focusing on the survival outcomes have demonstrated that the conservativeness of asymptotic tests without adjusting for stratification factors, as well as the inflated type I error rate of adjusted asymptotic tests conducted in a small sample of patients, can be relaxed using re-randomisation tests. Although several RCTs using minimisation have suggested the presence of non-proportional hazards (non-PH) effects, the application of re-randomisation tests has been limited to the log-rank test and Cox PH models, which may result in diminished statistical power when confronted with non-PH scenarios. To address this issue, we proposed two re-randomisation tests based on a maximum combination of weighted log-rank tests (MaxCombo test) and the difference in restricted mean survival time (dRMST) up to a fixed time point τ, both of which can be extended to adjust for randomisation stratification factors.

Methods We compared the performance of asymptotic and re-randomisation tests using the MaxCombo test, dRMST, log-rank test, and Cox PH models, assuming various non-PH situations for RCTs using minimisation, with total sample sizes of 50, 100, and 500 at a 1:1 allocation ratio. We mainly considered null, and alternative scenarios featuring delayed, crossing, and diminishing treatment effects.

Results Across all examined null scenarios, re-randomisation tests maintained the type I error rates at the nominal level. Conversely, unadjusted asymptotic tests indicated excessive conservatism, while adjusted asymptotic tests in both the Cox PH models and dRMST indicated inflated type I error rates for total sample sizes of 50. The stratified MaxCombo-based re-randomisation test consistently exhibited robust power across all examined scenarios.

Conclusions The re-randomisation test is a useful alternative in non-PH situations for RCTs with minimisation using the stratified MaxCombo test, suggesting its robust power in various scenarios.

Keywords Minimisation, Non-proportional hazards, Re-randomisation test, Weighted log-rank test, MaxCombo test, Restricted mean survival time, Survival analysis

*Correspondence:
Ryusei Kimura
kimura.ryusei@keio.jp
Full list of author information is available at the end of the article

Background

Randomisation has been widely used to evaluate the efficacy and safety of interventions in clinical trials, ensuring comparability by achieving the balance for treatment assignments across prognostic factors. In randomised controlled trials (RCTs) with limited sample sizes and several prognostic factors, simple randomisation may not be sufficient to balance treatment assignments across prognostic factors. In such cases, stratified randomisation or Pocock–Simon's minimisation method [1, 2] is often used. Stratified randomisation aims to balance the treatment assignments within each stratum; however, this objective becomes more challenging as the number of strata increases. Conversely, minimisation aims to achieve a marginal balance by sequentially assigning a new patient to the arm, which minimises the overall imbalance across the stratification factors. Consequently, RCTs that use minimisation are anticipated to have a higher number of strata relative to their sample sizes, necessitating careful consideration when choosing an analysis plan in these situations.

Previous studies have shown that a statistical test that relies on asymptotic normality without adjusting for the stratification factors used in minimisation is conservative [3–7]. Moreover, performing asymptotic tests with adjustment for all stratification factors may be unfeasible due to small or zero sample sizes within some strata. According to the FDA's covariate adjustment guideline [8], "sponsors should discuss their proposal with the relevant review division if the number of covariates is large relative to the sample size or if proposing to adjust for a covariate with many levels." In the survival analysis, adjusted Cox proportional hazards (PH) models exhibit an inflated type I error rate when the sample size is small [9]. Thus, regardless of the adjustment in the asymptotic test, the type I error rate may not be maintained at the nominal level. One potential solution is a re-randomisation test [10], which is expected to hold the type I error rate at the nominal level and improve power when the stratification factors are unadjusted in the test [11].

Several RCTs with minimisation have suggested the presence of non-PH treatment effects, which are typically classified as delayed, crossing, or diminishing. In an RCT with minimisation that compared the efficacy of pembrolizumab with that of a placebo in patients with completely resected stage III melanoma [12], non-PH effects were observed on recurrence-free survival. The results of this trial indicated a delayed effect in the overall population, where the survival probabilities of both groups remained similar for the first 3 months, and thereafter, the pembrolizumab group exhibited a higher survival probability than the placebo group. Moreover, a crossing effect was observed in a specific subgroup, where the pembrolizumab group initially demonstrated a lower survival probability than the placebo group; however, this trend was reversed in the later trial periods. Another study suggested a diminishing effect [13]. In these non-PH situations, the use of re-randomisation tests based on the log-rank test (LRT) and Cox PH models may reduce the statistical power, and the estimated hazard ratio (HR) may not be clinically interpretable [14, 15]. To the best of our knowledge, no previous study has evaluated the performance of re-randomisation tests under non-PH situations in RCTs using minimisation.

To address the limitations associated with non-PH situations, we proposed two re-randomisation tests based on statistics that do not rely on the PH assumption. First, we used a maximum combination of weighted LRT (WLRT) from the Fleming-Harrington (FH) $G^{\rho,\gamma}$ class [16], known as the MaxCombo test [17–20]. This test demonstrates a robust higher power compared to the LRT under some non-PH scenarios. The concept of using re-randomisation to derive the null distribution for such a maximum combination was described by Ganju et al. [21, 22]. They demonstrated that the unadjusted MaxCombo-based re-randomisation test, in a setting with simple randomisation, exhibits a robust high power nearly equivalent to the highest power achieved by the WLRT among combinations ($G^{0,0}$, $G^{1,0}$, $G^{0,1}$). However, they did not evaluate tests with different methods, such as dRMST, or stratified WLRTs and stratified MaxCombo tests. Second, we used a restricted mean survival time (RMST), defined as the mean survival time up to a fixed time point τ [23]. The difference in RMST (dRMST) can be clinically interpretable as follows, even under non-PH situations: "How long, on average, the time to event onset can be extended when patients are followed up until the specific time point τ." [24] These tests can be extended to adjust for randomisation stratification factors by stratification and regression adjustments.

In this study, we aimed to evaluate the performance of re-randomisation tests based on the aforementioned statistics, assuming various non-PH situations for RCTs with minimisation. We compared these methods in terms of their empirical type I error rate and power using numerical simulations. "Methods" section provides an overview of the testing procedure for both the asymptotic and re-randomisation tests. In "Simulation study" section, we explained the simulation settings and presented the results; the results are discussed in "Discussion" section.

Methods

We considered a two-armed comparison with a single survival primary endpoint in an RCT using minimisation that includes prognostic factors.

Testing null hypothesis

We defined the test statistics corresponding to the LRT and Cox PH model, and for the proposed methods, the MaxCombo test and dRMST, denoted as $Z^{(\mathrm{LRT})}$, $Z^{(\mathrm{Cox})}$, $Z_{\max}^{(\mathrm{WLRT})}$ and $Z^{(\mathrm{dRMST})}$, respectively. Furthermore, for these adjusted tests, we introduced $Z^{(\mathrm{SLRT})}$, $Z^{(\mathrm{ACox})}$, $Z_{\max}^{(\mathrm{SWLRT})}$, and $Z^{(\mathrm{IPCW})}$. All of these Z-based test statistics asymptotically follow a standard normal distribution under the null hypothesis. The comprehensive details of each statistic are provided in Appendices A to D.

Let $S_1(t)$ and $S_2(t)$ be the survival functions of the experimental and control arms, respectively. Throughout this manuscript, we focused on one-sided tests to demonstrate the superiority of our experimental arm. The tests, including the LRT, WLRT, and MaxCombo test, were based on the null hypothesis H_0 and alternative hypothesis H_1:

$$H_0 : S_1(t) = S_2(t) \text{ for all } t \geq 0,$$
$$H_1 : S_1(t) > S_2(t) \text{ for some } t \geq 0.$$

The H_0 was tested using $Z^{(\mathrm{LRT})}$, $Z_{\max}^{(\mathrm{WLRT})}$, or $Z_{\max}^{(\mathrm{SWLRT})}$. We defined the RMST up to time point τ for each arm $g\ (=1,2)$ as $\mu_g(\tau) = \int_0^\tau S_g(t)dt$. For dRMST, the hypotheses were as follows:

$$H_0 : \mu_1(\tau) - \mu_2(\tau) = 0 \text{ vs } H_1 : \mu_1(\tau) > \mu_2(\tau) = 0,$$

which was tested using $Z^{(\mathrm{dRMST})}$.

For the regression-based method, both the Cox and dRMST models were based on the following hypotheses:

$$H_0 : \beta_0 = 0 \text{ vs } H_1 : \beta_0 < 0,$$

where β_0 represents the coefficient parameter for the treatment effect. We can test H_0 using $Z^{(\mathrm{Cox})}$, $Z^{(\mathrm{ACox})}$ or $Z^{(\mathrm{IPCW})}$. The treatment effects β_0 were investigated rather than the other effects of the covariates.

Furthermore, we considered the strong null hypothesis [25, 26], which suggests that the survival probability in the experimental arm consistently remains less than that in the control arm, despite the hazard function initially favouring the control arm in the early trial periods:

$$H_0^{\mathrm{strong}} : S_1(t) \leq S_2(t) \text{ for all } t \geq 0,$$
$$H_1^{\mathrm{strong}} : S_1(t) > S_2(t) \text{ for some } t \geq 0.$$

Although the probability of falsely rejecting H_0^{strong} is expected to be below the nominal level, one-sided WLRTs from $G^{0,1}$, $G^{1,1}$, and the associated MaxCombo test may exhibit an inflated type I error rate in the strong null scenario without a covariate [26, 27]. This is because events early in the experimental arm unfairly favour this arm for these tests [28]. We evaluated the type I error rates of MaxCombo tests in strong null scenarios that incorporated prognostic factors.

Re-randomisation tests

When testing H_0 using a re-randomisation test, the treatment assignments are regenerated based on the actual randomisation procedure. During this regenerated treatment assignment process, the survival time, covariates, and order of patient entry remained fixed. Specifically, M datasets are generated to correspond with the observed dataset; these datasets include survival times and covariates identical to those in the observed dataset, along with regenerated treatment assignment sequences that may be identical to each other. For each iteration, we obtained the test statistics S_m for $m = 1, \ldots, M$ through Monte Carlo simulations. Subsequently, using the approximated null distribution derived from these iterations, the one-sided P-value for this approach was calculated using the formula $\sum_{m=1}^{M} I(S_m \geq S_{\mathrm{obs}})/M$, where S_{obs} represents the test statistic computed on the observed dataset [10].

Users can specify their preferred test statistics S_m and S_{obs} to test the corresponding H_0 as described in "Testing null hypothesis" section. The P-value for the MaxCombo-based asymptotic test is determined by the numerical integration of the multivariate normal distribution to account for multiplicity adjustment owing to the correlation among the four WLRT statistics (detailed in Appendix B). Conversely, for MaxCombo-based re-randomisation tests, $Z_{\max}^{(\mathrm{WLRT})}$ is directly adopted for both S_m and S_{obs}, thereby regenerating the approximated null distribution of $Z_{\max}^{(\mathrm{WLRT})}$. We can apply the stratified MaxCombo test, $Z_{\max}^{(\mathrm{SWLRT})}$, analogously. The null distribution for the maximum combination statistics, $Z_{\max}^{(\mathrm{WLRT})}$ and $Z_{\max}^{(\mathrm{SWLRT})}$, can be derived from the re-randomisation as explained by Ganju et al. [21, 22]. Re-randomisation tests based on other tests, such as the LRT, Cox PH models, and dRMST, can also be constructed using the corresponding Z-based test statistics.

A numerical issue occurs in which $Z^{(\mathrm{dRMST})}$ cannot be computed, due to censoring. This issue arises when the longest observed survival time in either arm is shorter than τ and is censored. Horiguchi et al. [29] illustrated this problem in detail and proposed several solutions. As their results indicated no differences between all evaluated methods, we simply adopted Method 2 [29], which extends the survival curve horizontally to τ. Although this extrapolation-based approach was originally employed for null distributions during re-randomisation and not for observed data, they regenerated the observed data until a pre-specified number of simulations were reached.

To reduce the simulation execution time, we applied Method 2 [29] to the observed data. Consequently, both $Z^{(\mathrm{dRMST})}$ and $Z^{(\mathrm{IPCW})}$ are computable, except when no events are observed in either arm. In these exceptional cases, neither $Z^{(\mathrm{dRMST})}$ nor $Z^{(\mathrm{IPCW})}$ can be computed owing to the failure to estimate their standard errors, even when using the method of Horiguchi et al.; thus, such cases were excluded from our simulation results.

Simulation study

Setup

To evaluate the performance of the aforementioned statistics in the asymptotic and re-randomisation tests, we calculated the empirical type I error rates and powers via numerical simulations, assuming two-armed RCTs with a 1:1 allocation ratio using minimisation. For the ith patient, the observed survival time was denoted as $T_i = \min(Y_i, C_i)$, where T_i denotes an event if $Y_i \leq C_i$, and otherwise, T_i denotes right censoring. We assumed that the censoring time C_i is independent of event time Y_i. Regarding prognostic factors, we set $Z_{i1}, Z_{i2} \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}(2/3)$ and $Z_{i3} \sim \mathrm{Bernoulli}(1/3)$. We generated Y_i following a piecewise exponential distribution with the rate parameter $\lambda_i(t)$, which is modelled as follows:

$$\lambda_i(t) = \lambda_0 \exp\left[\{\beta_1 I(t < \epsilon) + \beta_2 I(t \geq \epsilon)\} Z_{i0} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \gamma_3 Z_{i3}\right],$$

where $I(\cdot)$ denotes an indicator function, Z_{i0} is a treatment assignment based on minimisation with an assignment probability of 0.7. For all scenarios, we set covariate effects $\gamma_1, \gamma_2, \gamma_3$ as the common effect ($\gamma_1 = \gamma_2 = \gamma_3 = \gamma$), on the logarithm of the HR scale (log-HR). The treatment effects, β_1 and β_2, are on the log-HR scale and have been positioned before and after time point ϵ (months). The piecewise HR for treatment effects was different from the single HR from the Cox PH model. We assumed that a patient is uniformly accrued within 20 months and is followed up for at least 20 months; that is, C_i follows a uniform distribution on $[20, 40]$. The chosen total sample sizes, denoted by n, were 50, 100, and 500. All three prognostic factors were incorporated into the stratification using the minimisation scheme.

Initially, we considered two null scenarios: null scenarios with constant treatment effects (HR = 1.00 over time) and a strong null scenario A, where the experimental survival probability consistently favours that of the control arm. In strong scenario A, the HR for the treatment effect was $\exp(\beta_1) = 16.0$ in favour of the control arm for the initial month, subsequently shifting to $\exp(\beta_2) = 0.8$ in favour of the experimental arm. In this scenario, the one-sided WLRTs from $G^{1,1}$, $G^{0,1}$ and the associated MaxCombo test, may result in a false rejection and advocate the alternative hypothesis that supports the experimental arm [26]. The parameter settings were based on the studies by Freidlin et al. [25] and Roychoudhury et al. [27] and were slightly modified to adapt to our RCT setting with prognostic factors. Subsequently, our simulation results from the strong null scenario A deviated from our expectations, leading to the introduction of an additional strong null scenario B. This scenario exhibited weaker prognostic factor effects compared to scenario A, ensuring that the survival distribution for each stratum remained almost consistent with the marginal distribution. The survival plots for each stratum under strong null scenarios A and B are shown in Figures S1 and S2. The rationale behind the parameter settings for scenario B is described in "Empirical type I error rate" section. For alternative hypotheses, three non-PH scenarios were examined: delayed, diminishing, and crossing treatment effects. The marginal survival plots for each scenario are displayed in Fig. 1. The parameter settings for each scenario are presented in Table 1.

We evaluated the following tests: the MaxCombo test, with $(G^{0,0}, G^{1,0}, G^{1,1}, G^{0,1})$, and the dRMST with $\tau = 30$ (months). For comparative purposes, we also included the LRT and Wald tests based on the Cox PH models. In non-PH scenarios, using the Cox PH model, which is a model misspecification, may result in the loss of statistical power. The results of the other tests, including each WLRT from $G^{1,0}, G^{1,1}, G^{0,1}$, and the dRMST with $\tau = 20, 25$ (months), are presented in Table S1 and Figures S3 and S4. Furthermore, the performance of these tests was evaluated by adjusting the randomisation stratification factors using stratification and/or regression. We test H_0 at a one-sided significance level $\alpha = 0.025$. For re-randomisation tests with $M = 1,000$, we consistently used the Z-based test statistic for both S_m and S_{obs}. As described in "Re-randomisation tests" section, cases in which $Z^{(\mathrm{dRMST})}$ could not be computed due to censoring were addressed using Method 2 [29], which extends the survival curve horizontally to τ. Simulations were performed using the R software with the "survival" [30], "survRM2" [31], and "nph" [32] packages to obtain Z-based test statistics. The numbers of repetitions were 10,000 for the null scenarios and 5,000 for the other scenarios, including strong null scenarios.

Results

Empirical type I error rate

Null scenarios: Across all sample sizes ($n = 50, 100, 500$), the empirical type I error rates for the re-randomisation

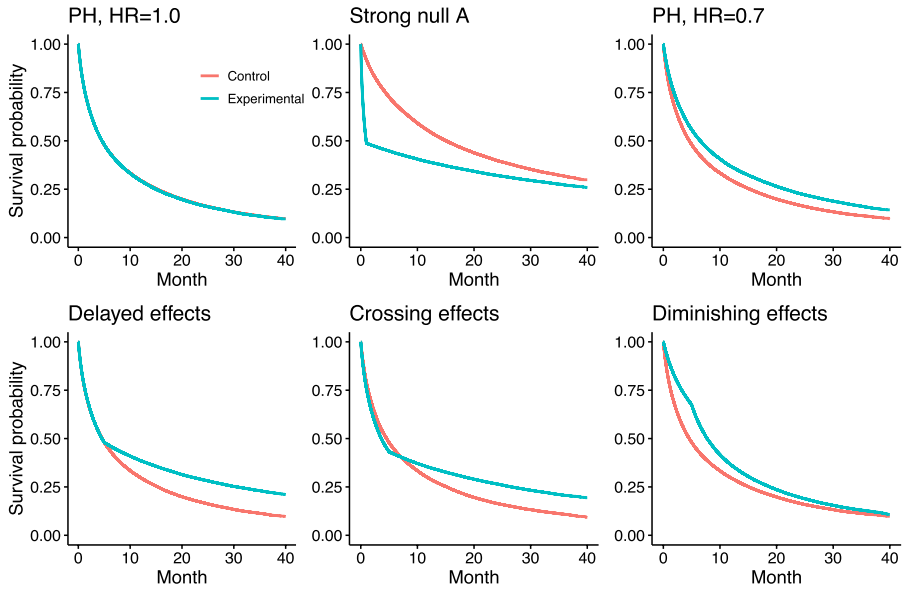

Fig. 1 The marginal survival plots for each scenario

Table 1 Parameter settings for each scenario

Scenario	λ_0	ϵ (month)	$\exp(\beta_1)$	$\exp(\beta_2)$	$\exp(\gamma)$	Censoring rates (%)	
						Test	Cont
Null	2.00	0.00	1.00	1.00	0.20	14	14
Strong null A	0.60	1.00	16.0	0.80	0.20	30	36
Strong null B	0.04	1.00	16.0	0.70	0.90	30	38
PH	2.00	0.00	0.70	0.70	0.20	19	14
Delayed	2.00	5.00	1.00	0.40	0.20	26	14
Crossing	2.00	5.00	1.25	0.40	0.20	24	14
Diminishing	2.00	5.00	0.40	0.95	0.20	16	14

tests were maintained at the nominal level 2.5% (Table 2). The unadjusted asymptotic tests exhibited conservative type I error rates for larger sample sizes. The adjusted asymptotic tests in the LRT and MaxCombo maintained a type I error rate of 2.5%, whereas those in Cox and dRMST showed inflated type I error rates when a limited sample size was used ($n = 50$). The detailed results of the other tests, including those of dRMST with different τ, are presented in Table S1.

Strong null scenarios: For strong null scenarios, we showed the type I error rates for MaxCombo tests in Table 3. The type I error rates of MaxCombo were not consistently at 0% across all sample sizes in both the asymptotic and re-randomisation tests, notably exceeding 2.5% at $n = 500$. However, the type I error rates of the stratified MaxCombo ranged from 0 to 0.12%.

Surprisingly, the trends in type I error rates for MaxCombo and stratified MaxCombo differed. This discrepancy between the MaxCombo and its stratified counterpart may be attributable to the differences in the calculation of the LR score, that is, regarding whether the calculation was performed marginally. Specifically, the stratum with $Z_{i1} = Z_{i2} = Z_{i3} = 0$ diverges from the marginal strong null scenario, as illustrated in Figure S1. To validate these observations, we further investigated the strong null scenario B. In this scenario, each stratum did not deviate substantially from the marginal settings by incorporating the weaker effects of the prognostic factors compared to that in scenario A (Figure S2). Both the MaxCombo and stratified MaxCombo showed inflated type I error rates (MaxCombo:1.90–4.80% and stratified MaxCombo:1.46–4.30%, Table 4). The results of the other tests in scenarios A and B are listed in Tables S2 and S3.

Table 2 Comparison of type I error rates under the null scenario

Method	Test	$n = 50$	$n = 100$	$n = 500$
		Type I error rate (%)		
Asymp	LRT	1.03	0.76	0.39
	MCT	1.41	0.96	0.66
	Cox	0.96	0.74	0.37
	dRMST	1.30	0.95	0.40
	Stratified LRT	2.68	2.35	2.35
	Stratified MCT	2.46	2.34	2.45
	Adjusted Cox	**3.03**	2.48	2.40
	Adjusted dRMST	**3.55**	2.80	2.59
Re-rand	LRT	2.62	2.49	2.51
	MCT	2.59	2.41	2.40
	Cox	2.62	2.49	2.51
	dRMST	2.59	2.51	2.45
	Stratified LRT	2.39	2.22	2.15
	Stratified MCT	2.23	2.22	2.49
	Adjusted Cox	2.67	2.27	2.28
	Adjusted dRMST	2.58	2.40	2.40

Abbreviations: *Asymp* Asymptotic test, *Re-rand* Re-randomisation test, *LRT* Log-rank test, *MCT* MaxCombo test, *dRMST* Difference in RMST, *n* total sample sizes

The range of Monte Carlo SE for type I error rates is 0.06 to 0.19 (%). Bold values exceed 2 × Monte Carlo SE + 2.50

Table 3 Comparison of type I error rates under the strong null scenario A

Method	Test	$n = 50$	$n = 100$	$n = 500$
		Type I error rate (%)		
Asymp	MCT	1.86	2.28	**8.28**
	Stratified MCT	0.10	0.04	0.00
Re-rand	MCT	**3.44**	**5.36**	**17.06**
	Stratified MCT	0.12	0.04	0.00

Abbreviations: Asymp Asymptotic test, *Re-rand* Re-randomisation test, *MCT* MaxCombo test, *n* total sample sizes

The maximum Monte Carlo SE for type I error rates (%) is 0.53. Bold values exceed 2 × Monte Carlo SE + 2.50

Table 4 Comparison of type I error rates under strong null scenario B

Method	Test	$n = 50$	$n = 100$	$n = 500$
		Type I error rate (%)		
Asymp	MCT	2.12	2.32	**4.80**
	Stratified MCT	1.66	1.78	**3.96**
Re-rand	MCT	1.90	2.14	**4.72**
	Stratified MCT	1.46	1.78	**4.30**

Abbreviations: Asymp Asymptotic test, *Re-rand* Re-randomisation test, *MCT* MaxCombo test, *n* total sample sizes

The range of Monte Carlo SE for type I error rates is 0.16 to 0.30 (%). Bold values exceed 2 × Monte Carlo SE + 2.50

Empirical power

Delayed effects scenarios: The unadjusted re-randomisation tests indicated higher statistical powers than their corresponding asymptotic tests across all sample sizes (Fig. 2 on the left). In particular, among all tests, the MaxCombo-based re-randomisation test exhibited the highest power. Moving to the adjusted tests, re-randomisation tests in the LRT and MaxCombo indicated powers similar to those of their corresponding asymptotic tests across all sample sizes. In contrast, the adjusted re-randomisation tests in the Cox PH models and dRMST indicated slightly lower powers than their corresponding asymptotic tests at $n = 50, 100$, with no substantial power differences at $n = 500$. Among these adjusted tests, the MaxCombo test exhibited the highest power at $n = 500$. For smaller sample sizes, $n = 50, 100$, almost no difference was observed among the adjusted tests, except for the asymptotic tests in the Cox PH models and dRMST.

The conservatism of the unadjusted asymptotic tests and the slightly higher power of the adjusted Cox PH models and dRMST at $n = 50, 100$ compared to their corresponding re-randomisation tests were consistently observed in subsequent scenarios. Therefore, we primarily focused on comparing the power of the re-randomisation tests in the following scenarios.

Crossing effects scenarios: Among the unadjusted re-randomisation tests, the MaxCombo-based re-randomisation test demonstrated a superior power, particularly at $n = 500$ (Fig. 2 at the center). Among the adjusted re-randomisation tests, both dRMST-based and MaxCombo-based re-randomisation tests exhibited higher powers than the other adjusted re-randomisation tests, especially at $n = 500$; no substantial power differences were observed among the adjusted re-randomisation tests at $n = 50, 100$. The power of dRMST depended on the value of τ (Figure S3).

Diminishing effect scenarios: Similar to the crossing effects scenarios, among the unadjusted re-randomisation tests, the MaxCombo-based re-randomisation test demonstrated a higher power (Fig. 2 on the right). Among the adjusted re-randomisation tests, the powers of the Cox/dRMST were the highest and lowest at $n = 50, 100$, respectively, with no substantial power differences at $n = 500$.

Proportional hazards scenarios: As supplementary information, the results, which include all evaluated tests for the PH scenario, are presented in Figure S4. No substantial power differences were observed among the unadjusted re-randomisation tests, except for WLRT with $G^{(0,1)}$. Conversely, among the adjusted re-randomisation tests, the Cox PH models demonstrated higher power than the other tests, consistent with the findings

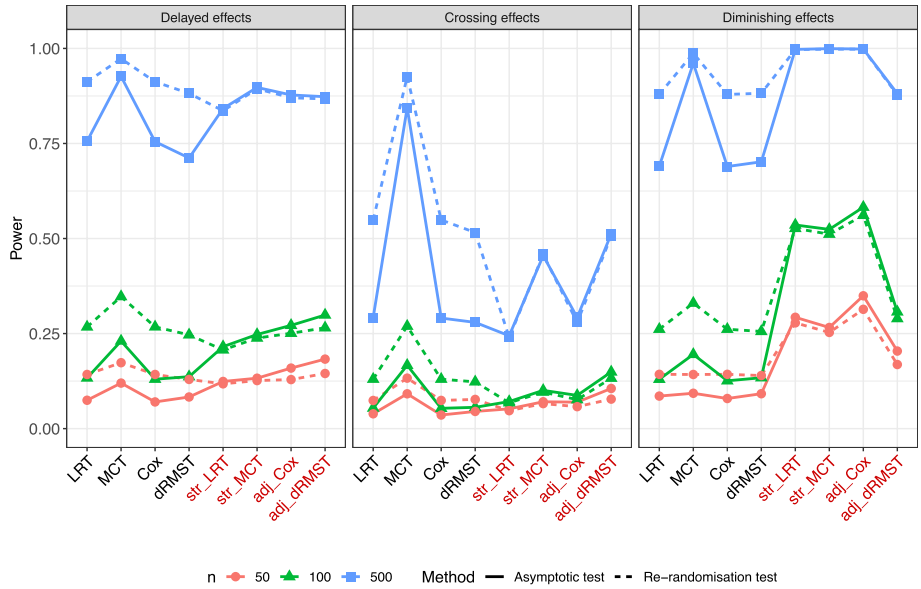

Fig. 2 Comparison of powers under three different non-PH scenarios. Abbreviations: LRT, log-rank; MCT, MaxCombo; dRMST, difference in RMST; str, stratification; adj, adjustment through a regression model; *n*, total sample sizes

reported by Xu et al. [9]. Following the Cox PH model, the LRT-based and MaxCombo-based re-randomisation tests showed the highest powers.

Discussion

Asymptotic tests versus re-randomisation tests

In the current study, we compared the performance of asymptotic and re-randomisation tests via numerical simulation, assuming various non-PH situations for RCTs with minimisation. As in previous studies [9, 33], the unadjusted asymptotic tests exhibited conservative type I error rates under null scenarios. Balancing treatment assignments by minimisation may lead to a correlation between treatment groups; thus, unadjusted tests that ignore this correlation yield conservative results. Accordingly, both the FDA guidelines and ICH E9 state the importance of accounting for the stratified randomisation factor in the analysis [8, 34]. Even with such adjustments, the Cox PH models and dRMST exhibited inflated type I error rates for $n = 50$ owing to the small sample size. This finding suggests that the asymptotic tests adjusted for stratification factors do not always yield valid results. Regardless of covariate adjustment, the type I error rates of the re-randomisation tests remained at the nominal level across all examined sample sizes. Except for cases with inflated type I error rates, the re-randomisation tests preserved almost the same power as their asymptotic counterparts. Therefore, a re-randomisation test is a valuable alternative to an asymptotic test in RCTs with minimisation.

The optimal re-randomisation test in terms of statistical power

Subsequently, we discussed which re-randomisation tests should be used, considering the aspect of power. Generally, the analysis used in RCTs must be pre-specified during the planning phase [34]. However, predicting the exact types of non-PH scenarios that may emerge is challenging, except in RCTs involving delayed effects in cancer immuno-oncology. Hence, a test that maintains robust power in various non-PH scenarios is required. Our simulation results show that the unadjusted re-randomisation tests in the LRT, Cox, and dRMST exhibited lower power than their corresponding adjusted re-randomisation tests. In contrast, the MaxCombo-based re-randomisation test demonstrated superior power under delayed and crossing effect scenarios, although it showed lower power than the other adjusted re-randomisation tests under diminishing effect scenarios. Among the adjusted re-randomisation tests, the LRT exhibited consistently lower powers across all examined non-PH scenarios. The adjusted Cox-based re-randomisation test showed the highest power under diminishing conditions, although it was inferior under the other non-PH scenarios. The adjusted dRMST-based re-randomisation test indicated a relatively high power for the crossing effect scenarios, but it was inferior in the other non-PH scenarios, with the power depending on τ. Finally, the stratified MaxCombo-based re-randomisation tests exhibited consistently superior powers across all examined scenarios, including the PH scenarios. This observation suggests that, in terms of power, the stratified

MaxCombo-based re-randomisation test is the optimal choice among examined tests.

Considerations for the stratified MaxCombo-based re-randomisation test in RCTs using minimisation

Despite the promising performance of the stratified MaxCombo-based re-randomisation test, its application in RCTs warrants caution due to potential inflation of the type I error rate in strong null scenarios. This concern, discussed in the literature [35–40], is particularly relevant when considering the strict accuracy requirements of primary analyses in RCTs. We note that these studies focused on RCTs using simple randomisation, excluding stratified MaxCombo tests. In contrast, we demonstrated that, even under minimisation, the type I error rates of the MaxCombo tests (alternatively, the re-randomisation test) were inflated, except in some scenarios. However, given our assumption of using it in a non-primary analysis context, the MaxCombo-based re-randomisation test remains an attractive option, while recognising its limitations in such extreme scenarios.

Furthermore, we examined two types of strong null scenarios A and B, and found that in the strong null scenario A, stratified MaxCombo tests exhibited lower type I error rates compared to MaxCombo tests. In both scenarios, the MaxCombo test exhibited a type I error rate of up to 17%. This result is consistent with the findings of a previous study [27]. In the strong null scenario A, the type I error inflation was not observed in the stratified MaxCombo test. This discrepancy arose because each stratum in scenario A deviated from the marginal setting (Figure S1 and S2). Under minimisation, several known prognostic factors with relatively strong effects exist, justifying the deviation of some strata from the marginal setting. Therefore, it is unrealistic to observe the inflation of the type I error rate for the stratified MaxCombo test under such an extreme null scenario in practical RCTs. This does not imply that the stratified MaxCombo test fundamentally overcomes this statistical flaw. However, even for those who do not accept the type I error rate inflation in the strong null scenario, it is worth considering the application of the stratified MaxCombo-based re-randomisation test as an option, possibly for non-primary analyses.

Interpretability of estimands in non-proportional hazards scenarios

In non-PH cases, the interpretability of the estimands corresponding to the selected test is important. In particular, the HR estimated using the Cox PH model may not be clinically interpretable [14, 15]. Moreover, the estimated HR in non-PH scenarios varied depending on the study-specific censoring time distribution, such as accrual rate, accrual period, and follow-up period. Therefore, the estimated HR cannot be interpreted as a simple or meaningful weighted average of the time-specific HR in non-PH scenarios [41, 42]. As an alternative, a piecewise HR, which describes the change in treatment effect over time, or a weighted HR corresponding to the maximum WLRT within the MaxCombo test, may be useful to capture the characteristics of the non-PH situation [20, 27, 43]. However, even these estimands may remain subject to criticism when interpreting the treatment effect causally [14, 44]. Shifting focus from HRs, the dRMST has a 1:1 correspondence between the testing and estimand, which is clinically interpretable even under non-PH scenarios. In crossing effects scenarios, where interpreting the HR becomes particularly challenging, the adjusted dRMST-based re-randomisation tests showed relatively high power. Although the power of dRMST depends on τ, considering that τ is typically selected from a clinical perspective, the dRMST-based re-randomisation test may be an attractive choice in terms of interpretability of the estimand. Importantly, in non-PH contexts, relying on a single summary measure and using test results alone to infer treatment efficacy is inadequate; thus, reporting multiple summary measures is recommended for a more comprehensive assessment of the treatment effect [27].

Conclusion

Re-randomisation tests have emerged as a credible and methodologically robust alternative to asymptotic tests in RCTs employing minimisation, particularly under non-PH conditions. The efficacy of the adjusted dRMST-based re-randomisation test is scenario-specific and significantly influenced by the strategic selection of the time point τ. In contrast, the stratified MaxCombo-based re-randomisation test has consistently demonstrated its pre-eminence in power across a broad spectrum of scenarios. Although whether an inflated type I error rate for the stratified MaxCombo-based re-randomisation test in the strong null scenario should be strictly controlled is debatable, this inflation is considerably reduced under some strong null scenarios for RCTs with minimisation. Consequently, considering the necessity of pre-specifying statistical analyses in RCT design, the stratified MaxCombo-based re-randomisation test is recommended for its steadfast and superior power in non-primary analysis.

Appendix

A Weighted log-rank test and its stratification

We initially explained the WLRT based on the FH $G^{\rho,\gamma}$ class [16]. Then, the stratified WLRT (SWLRT) can be naturally constructed by replacing the LR score for each stratum with the WLR score. Let t_j $(j = 1, 2, \ldots, D)$ be the jth ordered event time, d_{gj} be the number of events at t_j in arm g $(= 1, 2)$, and n_{gj} be the number of patients at risk of the event at t_j in arm g. The WLR score is obtained using the following equation:

$$U^{(\mathrm{WLRT})} = \sum_{j=1}^{D} w_j \left(d_{1j} - \frac{n_{1j} d_j}{n_j} \right),$$

where $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$. The variance of d_{1j} is given by:

$$v_{1j} = w_j^2 \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Subsequently, the variance of $U^{(\mathrm{WLRT})}$ is $\hat{V}\left[U^{(\mathrm{WLRT})}\right] = \sum_{j=1}^{D} v_{1j}$. Therefore, the statistic for WLRT is calculated as follows: $Z^{(\mathrm{WLRT})} = \frac{U^{(\mathrm{WLRT})}}{\sqrt{\hat{V}[U^{(\mathrm{WLRT})}]}}$, which asymptotically follows a standard normal distribution under the null hypothesis [45].

For the stratification analysis, we extended the indices to account for the stratum s $(= 1, 2, \ldots, S)$. Hence, n_{gjs} and d_{gjs} represent the number at risk and the number of events in stratum s, respectively. The WLR score for each stratum is calculated as follows:

$$U_s^{(\mathrm{WLRT})} = \sum_{j=1}^{D} w_j \left(d_{1js} - \frac{n_{1js} d_{js}}{n_{js}} \right),$$

where $d_{js} = d_{1js} + d_{2js}$ and $n_{js} = n_{1js} + n_{2js}$. The variance of d_{1js} is calculated as follows:

$$v_{1js} = w_j^2 \frac{n_{1js} n_{2js} d_{js} (n_{js} - d_{js})}{n_{js}^2 (n_{js} - 1)}.$$

Subsequently, the variance of $U_s^{(\mathrm{WLRT})}$ is calculated as follows: $\hat{V}\left[U_s^{(\mathrm{WLRT})}\right] = \sum_{j=1}^{D} v_{1js}$. Therefore, the test statistic for SWLRT is defined as follows:

$$Z^{(\mathrm{SWLRT})} = \frac{\sum_{s=1}^{S} U_s^{(\mathrm{WLRT})}}{\sqrt{\sum_{s=1}^{S} \hat{V}\left[U_s^{(\mathrm{WLRT})}\right]}},$$

which asymptotically follow a standard normal distribution under the null hypothesis [45].

The weight function for the $G^{\rho,\gamma}$ class is defined as $w_j = \left(\hat{S}(t_{j-1})\right)^{\rho} \left(1 - \hat{S}(t_{j-1})\right)^{\gamma}$, where $\rho \geq 0, \gamma \geq 0$, and $\hat{S}(t_{j-1})$ denote the pooled KM estimator. The representative FH classes, $G^{1,0}, G^{1,1}$, and $G^{0,1}$ are weighted to the early, middle, and late events, respectively, whereas $G^{0,0}$ corresponds to the LRT. The WLRT with unfavourable weights that do not align with the observed data may exhibit a reduced power [20]. If the non-PH type, such as delayed, crossing, or diminishing treatment effects, is unpredictable, a robust method for non-PH types may be desirable.

B MaxCombo test and its stratification

A maximum combination of WLRT statistics, known as the MaxCombo test, demonstrates a robust higher power compared to LRT under a non-PH situation [17–20]. The MaxCombo for one-sided test is defined as follows:

$$Z_{\max}^{(\mathrm{WLRT})} = \max\left(Z_1^{(\mathrm{WLRT})}, Z_2^{(\mathrm{WLRT})}, \ldots, Z_L^{(\mathrm{WLRT})}\right),$$

where $Z_l^{(\mathrm{WLRT})}$ $(l = 1, 2, \ldots, L)$ denotes some classes of the WLRT. Among several combinations of $Z_l^{(\mathrm{WLRT})}$, Lin et al. [20] proposed the MaxCombo test based on $Z_{\max}^{(\mathrm{WLRT})}$ with $Z_1^{(\mathrm{WLRT})}, Z_2^{(\mathrm{WLRT})}, Z_3^{(\mathrm{WLRT})}$, and $Z_4^{(\mathrm{WLRT})}$ corresponding to $G^{0,0}, G^{1,0}, G^{1,1}$, and $G^{0,1}$. The $Z_{\max}^{(\mathrm{WLRT})}$ asymptotically follows a multivariate normal distribution with means of zero and the correlation matrix $\boldsymbol{R}$ under the null hypothesis [19, 20]. Let the covariances of G^{ρ_i,γ_i} and G^{ρ_j,γ_j} be $\mathrm{Cov}[G^{\rho_i,\gamma_i}, G^{\rho_j,\gamma_j}]$ and let the (i, j)-component r_{ij} of $\boldsymbol{R}$ be given by

$$r_{ij} = \begin{cases} \frac{\mathrm{Cov}[G^{\rho_i,\gamma_i}, G^{\rho_j,\gamma_j}]}{\sqrt{V[G^{\rho_i,\gamma_i}]V[G^{\rho_j,\gamma_j}]}} = \frac{V\left[G^{\frac{\rho_i+\rho_j}{2},\frac{\gamma_i+\gamma_j}{2}}\right]}{\sqrt{V[G^{\rho_i,\gamma_i}]V[G^{\rho_j,\gamma_j}]}} & \text{for } i \neq j \\ 1 & \text{for } i = j, \end{cases}$$

as described previously [19, 20, 27]. The P-value of the MaxCombo test was calculated by numerical integration based on the above multivariate normal distribution with using the algorithm proposed by Genz [46]. For the adjustment of covariates, the stratified MaxCombo test statistics, $Z_{\max}^{(\mathrm{SWLRT})}$ can be naturally constructed by replacing $Z_l^{(\mathrm{WLRT})}$ with $Z_l^{(\mathrm{SWLRT})}$.

C Cox proportional hazards model

The Cox PH model [47] is routinely used to estimate the HR as a treatment effect under the PH assumption. Let $\boldsymbol{Z}_i$ be a $p + 1$-dimensional covariate vector comprising a treatment group indicator and p covariates. For multivariable analysis, the Cox PH model is expressed as follows:

$$h(\boldsymbol{Z}_i, t) = \exp(\boldsymbol{\beta}^{\top} \boldsymbol{Z}_i) h_0(t),$$

where $h(\boldsymbol{Z}_i, t)$ represents the hazard function for the ith patient at time t; vector $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_p)$ is a $p+1$ -dimensional regression coefficients, including a treatment effect β_0; and $h_0(t)$ is a baseline hazard function for time t.

Moreover, the stratified Cox PH model is expressed using the following equation:

$$h_s(Z_i, t) = \exp(\beta_0^* Z_i) h_{0s}(t),$$

where $h_s(Z_i, t)$ represents the hazard function for ith patient at time t in stratum s $(= 1, \ldots, S)$, Z_i is the treatment group indicator for ith patient, β_0^* is the coefficient for the treatment group indicator, and $h_{0s}(t)$ is the baseline hazard function for time t in stratum s.

The estimation of $\boldsymbol{\beta}$ and β_0^* are based on a partial likelihood [45]. Given our interest in treatment effects, Wald tests based on the aforementioned equations are expressed as follows: $Z^{(\mathrm{ACox})} = \frac{\hat{\beta}_0}{\sqrt{\hat{V}[\hat{\beta}_0]}}$ and $Z^{(\mathrm{SCox})} = \frac{\hat{\beta}_0^*}{\sqrt{\hat{V}[\hat{\beta}_0^*]}}$, respectively. For univariate situations with $p = 0$, we denote the test statistics by $Z^{(\mathrm{Cox})}$. These test statistics asymptotically follow a standard normal distribution under the null hypothesis [45].

D Difference in RMST and its adjustment

The RMST is defined as the mean survival time up to a fixed time point τ [23]. Let $D_g(\tau)$ be the total number of events up to time point τ in arm g $(= 1, 2)$ and $\hat{S}_g(t)$ be a KM estimator of $S_g(t)$. For each arm, the RMST estimated by the direct integration of the KM curve method [45, 48] is calculated as follows:

$$\hat{\mu}_g(\tau) = \int_0^\tau \hat{S}_g(t)dt = \sum_{j=0}^{D_g(\tau)} (t_{j+1} - t_j)\hat{S}_g(t_j),$$

where $t_0 = 0$ and $t_{D_g(\tau)+1} = \tau$. The variance of $\hat{\mu}_g(\tau)$ based on Greenwood's formula [49] is calculated as follows:

$$\hat{V}[\hat{\mu}_g(\tau)] = \sum_{j=1}^{D_g(\tau)} \left[\sum_{k=j}^{D_g(\tau)} (t_{k+1} - t_k)\hat{S}_g(t_k) \right]^2 \frac{d_{gj}}{n_{gj}(n_{gj} - d_{gj})}.$$

Thus, the test statistics for the dRMST were constructed as

$$Z^{(\mathrm{dRMST})} = \frac{\hat{\mu}_1(\tau) - \hat{\mu}_2(\tau)}{\sqrt{\hat{V}[\hat{\mu}_1(\tau)] + \hat{V}[\hat{\mu}_2(\tau)]}},$$

which asymptotically follows the standard normal distribution under the null hypothesis [48].

For covariate adjustments, the inverse probability of censoring weighted (IPCW) method was proposed by Tian et al. [50]. Let C_i $(i = 1, \ldots, n)$ be a non-negative random variable denoting the censoring time for the ith patient, and $\eta(\cdot)$ be a general link function. The covariate vector $\boldsymbol{\beta}$, which includes a treatment effect β_0, can be estimated using the estimating equation for the IPCW, based on $\eta(E[Y_i(\tau) \mid \boldsymbol{X}_i]) = \boldsymbol{\beta}^\top \boldsymbol{X}_i$, where $\boldsymbol{X}_i^\top = (1, \boldsymbol{Z}_i^\top)$. The estimating equation for the IPCW is obtained as follows:

$$U^{(\mathrm{IPCW})}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{j=1}^{D} \frac{1}{\hat{G}_{g_j}(t_j(\tau))} \boldsymbol{X}_j \{t_j(\tau) - \eta^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_j)\},$$

where $\hat{G}_g(\cdot)$ represents the KM estimators of C_i in arm g_j $(= 1, 0)$ for the patient who experienced the jth event, and $t_j(\tau) = \min(t_j, \tau)$ denotes the restricted time to event. The variance of $\hat{\boldsymbol{\beta}}$ can be estimated using a sandwich estimator. Hence, an IPCW-based test statistic is constructed using $Z^{(\mathrm{IPCW})} = \frac{\hat{\beta}_0}{\sqrt{\hat{V}(\hat{\beta}_0)}}$, which asymptotically follows a standard normal distribution under the null hypothesis.

Abbreviations

RCT	Randomised controlled trial
PH	Proportional hazards
KM	Kaplan–Meier
FH	Fleming–Harrington
LRT	Log-rank test
WLRT	Weighted log-rank test
SWLRT	Stratified weighted log-rank test
dRMST	difference in restricted mean survival time
IPCW	Inverse probability of censoring weighted

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02295-2.

Additional file 1. Survival plots for each stratum under strong null scenarios are illustrated in Figure S1 and S2; simulation results under the null scenarios are provided in Table S1; results under the strong null scenarios for other tests not shown in the manuscript are provided in Tables S2 and S3; results under the non-PH scenarios for other tests not shown in the manuscript are illustrated in Figure S3; results under the PH scenarios are provided in Figure S4.

Acknowledgements

We would like to thank Editage (https://www.editage.com) for providing excellent English language editing assistance.

Authors' contributions

RK performed the simulation studies and prepared the manuscript. SN, KN, and YS interpreted the results and revised the manuscript. All authors have read and approved the final version of the manuscript.

Funding

This study was supported by JST SPRING (Grant Number: JPMJSP2123) and AMED (Grant Numbers: JP23lk1503008 and JP23lk0201701). The funding bodies played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Availability of data and materials
The datasets used and/or analysed in the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Biostatistics Unit, Clinical and Translational Research Center, Keio University Hospital, Tokyo 160-8582, Japan. [2]Graduate School of Health Management, Keio University, Tokyo 252-0822, Japan. [3]Department of Biostatistics and Bioinformatics, Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan. [4]Department of Preventive Medicine and Public Health, Keio University School of Medicine, Tokyo 160-8582, Japan.

Received: 16 November 2023 Accepted: 24 July 2024
Published online: 30 July 2024

References

1. Taves DR. Minimization: A new method of assigning patients to treatment and control groups. Clin Pharmacol Ther. 1974;15(5):443–53. https://doi.org/10.1002/cpt1974155443.
2. Pocock SJ, Simon R. Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. Biometrics. 1975;31(1):103–15.
3. Shao J, Yu X, Zhong B. A theory for testing hypotheses under covariate-adaptive randomization. Biometrika. 2010;97(2):347–60.
4. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. Stat Med. 2012;31(4):328–40.
5. Ma W, Hu F, Zhang L. Testing hypotheses of covariate-adaptive randomized clinical trials. J Am Stat Assoc. 2015;110(510):669–80.
6. Ma W, Qin Y, Li Y, Hu F. Statistical inference for covariate-adaptive randomization procedures. J Am Stat Assoc. 2020;115(531):1488–97.
7. Ye T, Shao J. Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. J R Stat Soc Ser B Stat Methodol. 2020;82(5):1301–23.
8. FDA. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. 2021. https://www.fda.gov/media/148910/download. Accessed 27 July 2024.
9. Xu Z, Proschan M, Lee S. Validity and power considerations on hypothesis testing under minimization. Stat Med. 2016;35(14):2315–27. https://doi.org/10.1002/sim.6874.
10. Rosenberger WF, Lachin JM. Randomization in clinical trials: theory and practice. Wiley, 2015.
11. Simon R, Simon NR. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. Stat Probab Lett. 2011;81(7):767–72.
12. Eggermont AM, Blank CU, Mandala M, Long GV, Atkinson V, Dalle S, et al. Adjuvant pembrolizumab versus placebo in resected stage III melanoma. N Engl J Med. 2018;378(19):1789–801.
13. Mehanna H, Wong WL, McConkey CC, Rahman JK, Robinson M, Hartley AG, et al. PET-CT surveillance versus neck dissection in advanced head and neck cancer. N Engl J Med. 2016;374(15):1444–54.
14. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13–15.
15. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014;32(22):2380–5.
16. Fleming TR, Harrington DP. Counting processes and survival analysis. John Wiley & Sons; 2013.
17. Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. Biometrics. 1996;52(8):721–5.
18. Lee SH. On the versatility of the combination of the weighted log-rank statistics. Comput Stat Data Anal. 2007;51(12):6557–64.
19. Karrison TG. Versatile tests for comparing survival curves based on weighted log-rank statistics. Stata J. 2016;16(3):678–90.
20. Lin RS, Lin J, Roychoudhury S, Anderson KM, Hu T, Huang B, et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. Stat Biopharm Res. 2020;12(2):187–98.
21. Ganju J, Yu X, Ma G. Robust inference from multiple test statistics via permutations: a better alternative to the single test statistic approach for randomized trials. Pharm Stat. 2013;12(5):282–90.
22. Ganju J, Ma G. The potential for increased power from combining P-values testing the same hypothesis. Stat Methods Med Res. 2017;26(1):64–74.
23. Irwin J. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. Epidemiol Infect. 1949;47(2):188–9.
24. Hasegawa T, Misawa S, Nakagawa S, Tanaka S, Tanase T, Ugai H, et al. Restricted mean survival time as a summary measure of time-to-event outcome. Pharm Stat. 2020;19(4):436–53.
25. Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? J Clin Oncol. 2019;37(35):3455–9.
26. Magirr D, Burman CF. Modestly weighted logrank tests. Stat Med. 2019;38(20):3782–90.
27. Roychoudhury S, Anderson KM, Ye J, Mukhopadhyay P. Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. Stat Biopharm Res. 2021;15(2):280–94.
28. Magirr D. Non-proportional hazards in immuno-oncology: Is an old perspective needed? Pharm Stat. 2021;20(3):512–27.
29. Horiguchi M, Uno H. On permutation tests for comparing restricted mean survival time with small sample from randomized trials. Stat Med. 2020;39(20):2655–70.
30. Therneau TM. A Package for Survival Analysis in R. 2022. R package version 3.4-0. https://CRAN.R-project.org/package=survival. Accessed 13 Sept 2023.
31. Uno H, Tian L, Horiguchi M, Cronin A, Battioui C, Bell J. survRM2: Comparing Restricted Mean Survival Time. 2022. R package version 1.0-4. https://CRAN.R-project.org/package=survRM2. Accessed 13 Sept 2023.
32. Ristl R, Ballarini NM, Götte H, Schüler A, Posch M, König F. Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology. Pharm Stat. 2021;20(1):129–45.
33. Hagino A, Hamada C, Yoshimura I, Ohashi Y, Sakamoto J, Nakazato H. Statistical comparison of random allocation methods in cancer clinical trials. Control Clin Trials. 2004;25(6):572–84.
34. ICH E9. Statistical principles for clinical trials. 1998. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Accessed 27 July 2024.
35. Posch M, Ristl R, König F. Testing and Interpreting the "Right" Hypothesis—Comment on "Non-proportional Hazards—An Evaluation of the MaxCombo Test in Cancer Clinical Trials". Stat Biopharm Res. 2022;15(2):310–1.
36. Shen YL, Wang X, Sirisha M, Mulkey F, Zhou J, Gao X, et al. Nonproportional Hazards—An Evaluation of the MaxCombo Test in Cancer Clinical Trials. Stat Biopharm Res. 2023;15(2):300–9.
37. Lin RS, Mukhopadhyay P, Roychoudhury S, Anderson KM, Hu T, Huang B, et al. Comment on "Non-Proportional Hazards-an Evaluation of the MaxCombo Test in Cancer Clinical Trials" by the Cross-Pharma Non-Proportional Hazards Working Group. Stat Biopharm Res. 2023;15(2):312–4.
38. Shen YL, Mushti S, Mulkey F, Gwise T, Wang X, Zhou J, et al. Rejoinder to Comments on "Non-Proportional Hazards-An Evaluation of the MaxCombo Test in Cancer Clinical Trials." Stat Biopharm Res. 2023;15(2):315–7.
39. Magirr D, Burman CF. The strong null hypothesis and the maxcombo test: Comment on "robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance form a cross-pharma working group.". Stat Biopharm Res. 2023;15(2):295–6.
40. Magirr D, Burman CF. The MaxCombo Test Severely Violates the Type I Error Rate. JAMA Oncol. 2023;9(4):571–2.

41. Horiguchi M, Hassett MJ, Uno H. How do the accrual pattern and follow-up duration affect the hazard ratio estimate when the proportional hazards assumption is violated? Oncologist. 2019;24(7):867–71.
42. Uno H, Horiguchi M. Ratio and difference of average hazard with survival weight: new measures to quantify survival benefit of new therapy. Stat Med. 2023;42(7):936–52.
43. Lin RS, León LF. Estimation of treatment effects in weighted log-rank tests. Contemp Clin Trials Commun. 2017;8:147–55.
44. Bartlett JW, Morris TP, Stensrud MJ, Daniel RM, Vansteelandt SK, Burman CF. The hazards of period specific and weighted hazard ratios. Stat Biopharm Res. 2020;12(4):518–9.
45. Collett D. Modelling survival data in medical research. CRC Press; 2015.
46. Genz A. Numerical computation of multivariate normal probabilities. J Comput Graph Stat. 1992;1(2):141–9.
47. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–202.
48. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(1):1–15.
49. Greenwood M, et al. A report on the natural duration of cancer. Reports on Public Health and Medical Subjects: Her Majesty's Stationery Office. 1926;(33).
50. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2013;15(2):222–33. https://doi.org/10.1093/biostatistics/kxt050.

Publisher's Note