Journal of Data Science 23 (3), 491–498
July 2025

DOI: 10.6339/25-JDS1177
Statistical Data Science

Restricted Mean Survival Time for a Randomized Study with Survival Outcome

Guogen Shan[1],*
[1]*Department of Biostatistics, University of Florida, Gainesville FL, 32610, USA*

Abstract

When comparing two survival curves, three tests are widely used: the Cox proportional hazards test, the logrank test, and the Wilcoxon test. Despite their popularity in survival data analysis, there is no clear clinical interpretation especially when the proportional hazard assumption is not valid. Meanwhile, the restricted mean survival time (RMST) offers an intuitive and clinically meaningful interpretation. We compare these four tests with regards to statistical power under many configurations (e.g., proportional hazard, early benefit, delayed benefit, and crossing survivals) with data simulated from the Weibull distributions. We then use an example from a lung cancer trial to compare their required sample sizes. As expected, the CoxPH test is more powerful than others when the PH assumption is valid. The Wilcoxon test is often preferable when there is a decreasing trajectory in the event rate as time goes. The RMST test is much more powerful than others when a new treatment has early benefit. The recommended test(s) under each configuration are suggested in this article.

Keywords *Cox proportional hazards model; logrank test; randomized trial; restricted mean survival time; Wilcoxon test*

1 Introduction

Restricted mean survival time (RMST) is a summary measure for a study with survival outcome: an average treatment effectiveness over a pre-specified time period (Uno et al., 2015). It is computed as the area under the estimated survival curve (e.g., the Kaplan-Meier curve) from baseline to the clinically meaningful follow-up time (τ). The RMST offers an easily understandable interpretation as the expected survival time within a given time window from time 0 to τ (Liao et al., 2020). The upper limit of the time window, τ, could affect the statistical power. As pointed out by Tian et al. (2018), the pre-specified time τ in computing the RMST should be no larger than the largest observed time from the study to provide valid statistical inference.

In a parallel randomized clinical trial with survival outcome, the difference in the RMST between a new treatment and the standard treatment can be estimated by using its asymptotic limiting normal distribution or simulation studies (Tian et al., 2018). When the proportional hazard (PH) assumption is met for comparing two survival curves, the Cox PH (CoxPH) model is commonly used. In the non-PH scenarios, nonparametric tests are widely used: the logrank test and the Wilcoxon test (Harrington and Fleming, 1982; Shan, 2023). The logrank test is computed by using the difference of the observed score and the expected score similar to a chi-squared test. The Wilcoxon test by Peto and Peto (1972) a weighted version of the logrank test,

* Email: gshan@ufl.edu.

Received June 21, 2024; Accepted March 17, 2025

and it can also be considered as an extension of the Wilcoxon rank sum test for two independent samples in the presence of survival outcome.

When the PH assumption is met, the CoxPH test is always more powerful than others (Tian et al., 2018; Zhang et al., 2024; Shan, 2022). The CoxPH test has higher statistical power than the RMST test when a new treatment has delayed benefit as compared to the standard treatment (Tian et al., 2018). Royston et al. (2013) compared the RMST test and the logrank test under the PH and the non-PH scenarios. The logrank test is more powerful than the RMST for studies that meet the PH assumption, while the trend is reversed for the non-PH scenarios (Shan, 2022; Shan et al., 2025). In this article, we conduct extensive simulation studies to compare the performance of the RMST test and the commonly used tests with regards to statistical power under many scenarios to fill the gap with no comprehensive comparisons among these tests.

The rest of the article is organized as follows. In Section 2, we introduce the four tests for comparing two survival curves. Then, we compare their performance with regards to statistical power in Section 3. At the end of that section, we use an example from a lung cancer trial to illustrate the application of the four tests in sample size calculation in practice. Lastly, we provide some comments in Section 4.

2 Methods

In a study to compare a new treatment and the standard treatment with survival outcome as the primary endpoint or the secondary endpoint, we are often interested in testing whether the new treatment has a higher survival rate than the standard treatment, with the hypotheses given as:

$$H_0 : S_1(t) \leqslant S_0(t) \quad \text{against} \quad H_a : S_1(t) > S_0(t), \tag{1}$$

where S_0 and S_1 are the survival functions of the standard treatment and the new treatment, respectively. In practice, the total study time (t_T) is often approximately pre-specified due to study timeline and/or budget, and it can be split into two parts: patient accrual time (t_a) and the follow-up time (t_f), with $t_T = t_a + t_f$.

Multiple tests have been proposed to test the aforementioned hypotheses in Equation (1). The first two widely used tests are: the logrank test (Mantel, 1966), and the Wilcoxon test (Peto and Peto, 1972). Both tests are nonparametric tests, and their difference lies in the weight function, $[S(t)]^\rho$, in the test statistic (Harrington and Fleming, 1982): $\rho = 0$ for the logrank test and $\rho = 1$ for the Wilcoxon test. These two tests can be implemented by using the R function $survdiff$ with the value of $rho = 0$ and 1. The third test is the widely used CoxPH test which assumes the PH assumption (Cox, 1972). This test can be computed by using the R function $coxph$. It should be noted that the logrank test and the Wilcoxon test are based on a chi-squared test for a two-sided hypothesis, and we add a sign to these two tests with the difference between the observed events and expected events from the new treatment group for the one-sided hypothesis in Equation (1).

In addition to the commonly used three tests, the RMST is an alternative test that could be used to compare the area under the survival curve from time 0 to a pre-specified time τ, specifically,

$$\mu(\tau) = \int_0^\tau S(t)dt. \tag{2}$$

Then, the hypotheses to test the effectiveness of a new treatment as compared to the standard treatment are written as

$$H_0 : \mu_1(\tau) \leqslant \mu_0(\tau) \quad \text{against} \quad H_a : \mu_1(\tau) > \mu_0(\tau),$$

where $\mu_1(\tau)$ and $\mu_0(\tau)$ are the RMST for the new treatment and the RMST for the standard treatment, respectively. The RMST test statistic comparing two groups can be computed by using the R function $rmst2$ from the $survRM2$ package (Uno, 2017). We compare the performance of these four tests with regards to statistical power in the next section with extensive simulation studies.

3 Results

We compare the performance of the RMST, the logrank test, the CoxPH test, and the Wilcoxon test with regards to statistical power under the PH assumption and the non-PH assumption. The Weibull distribution is used in simulation studies, with the survival function as:

$$S(t) = e^{-(\frac{t}{\lambda})^k},$$

where k is the shape parameter and λ is the scale parameter. In the Weibull distribution, the two parameters can be specified to make different scenarios of the survival curve of the new treatment as compared to that of the standard treatment. The null data are simulated from Weibull(k_0, λ_0), and data under the alternative are simulated from Weibull(k_1, λ_1). In the following simulation studies, we assume that the patient accrual time is $t_a = 12$ months, and the follow-up time is $t_f = 24$ months. In the simulation studies, the total time for RMST is assumed to be the same as the follow-up time: $\tau = t_f = 24$ months. To have a fair comparison between these methods, we first determine the threshold value for the test statistic under the null hypothesis to have the nominal level of 5% (one-sided). Then, that threshold value is used in the simulated statistical power calculation for each method.

When $k = 1$, the Weibull distribution becomes the exponential distribution. In Figure 1, we consider the Weibull distribution with $k_0 = 1$ and $\lambda_0 = 25$ for the null data, and the alternative distributions with $\lambda_1 = 25, 27$, and 35, and k_1 from 0.6 to 3.5. The survival functions under the null and those under the alternative with $k_1 = 2$ are presented in the first column of the figure. It can be seen that when λ_1 is close to λ_0, the new treatment appears to have early treatment benefit. In the configurations with early benefit, the RMST test has the highest power, followed by the Wilcoxon test, and the other two tests. The power gain using the RMST test is sizable when k_1 is much larger than k_0. When $k_1 \leqslant k_0$, the CoxPH test and the logrank test are slightly more powerful than the other two test in many cases. Given k_1, the power difference among the four tests gets smaller as λ_1 goes up.

When the null data suggests a decreasing trajectory in the event rate as time goes, the shape parameter under null is less than 1: $k_0 < 1$. We present the power comparison among the four tests when $k_0 = 0.2, 0.5$, and 0.8 in Figure 2. When k_0 is very small (e.g., 0.2) and $k_1 > k_0$, the Wilcoxon test is much more powerful than the other three tests. As k_0 is increased to 0.5, the Wilcoxon test still has the highest power, followed by the RMST test, the CoxPH test, and the logrank test. The CoxPH test and the logrank test often have similar statistical power, although the CoxPH test could be slightly more powerful in general. As k_0 is close to 1 (e.g., $k_0 = 0.8$), the findings are similar to these in Figure 1: the RMST test becomes the one having the highest statistical power in many configurations.

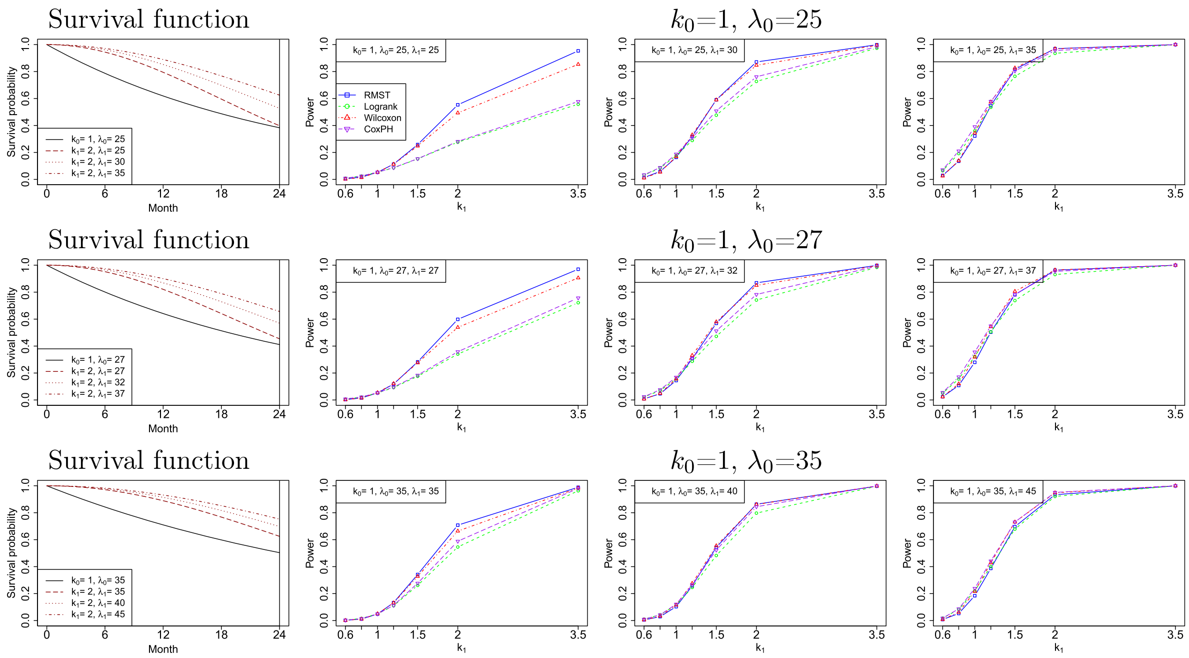

Figure 1: Power comparisons among the four tests when $k_0 = 1$ in the Weibull distribution. The first column shows the survival curve under the null hypothesis and the selected curves under the alternative hypotheses. The other three columns are for the computed power values as a function of k_1 when λ_1 is fixed.

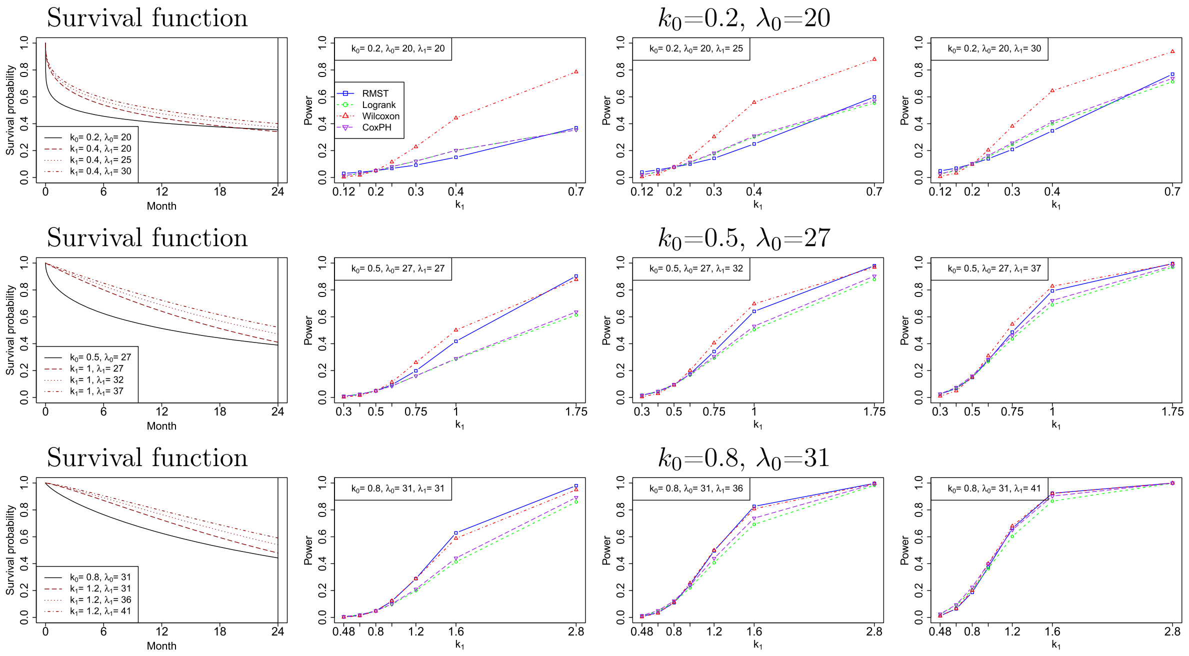

Figure 2: Power comparisons among the four tests when $k_0 < 1$ in the Weibull distribution. The first column shows the survival curve under the null hypothesis and the selected curves under the alternative hypotheses. The other three columns are for the computed power values as a function of k_1 when λ_1 is fixed.

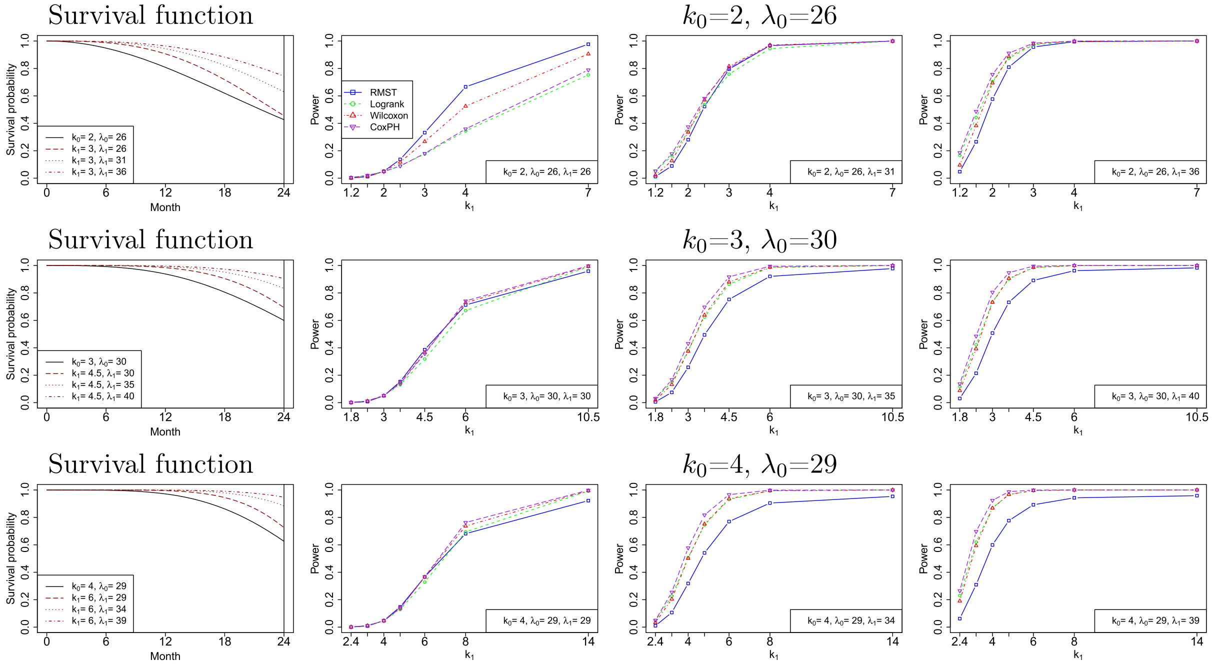

Figure 3: Power comparisons among the four tests when $k_0 > 1$ in the Weibull distribution. The first column shows the survival curve under the null hypothesis and the selected curves under the alternative hypotheses. The other three columns are for the computed power values as a function of k_1 when λ_1 is fixed.

It is common that the disease event rate increases as time goes. In such cases, k_0 is larger than 1. Figure 3 shows the power of the four tests when $k_0 = 2, 3$, and 4. In the majority of the configurations, the CoxPH test becomes the most powerful test, followed by the Wilcoxon test, the logrank test, and the RMST test. We found that the Wilcoxon test and the logrank test have similar power, and the CoxPH test has substantial power gain as compared to these two tests. The power gains of these three tests over the RMST test goes up as k_0 increases. In the cases with a small k_0, the RMST test could still be preferable when λ_1 is close to λ_0.

3.1 Example

We use a non-small cell lung cancer (NSCLC) trial as an example to illustrate the application of the four tests in sample size calculation for a parallel randomized trial for patients with refractory or recurrent NSCLC (Takiguchi et al., 2007). That trial was a single-arm study with all patients treated by a combination of irinotecan and cisplatin. Suppose an investigator is going to conduct a parallel study to compare a new treatment with that treatment with survival outcome as the primary endpoint. This is a randomized trial with an equal sample size in each group with $t_a = 12$ months and $t_f = 18$ months. We assume the survival function follows the Weibull distribution. From the presented survival curve (Takiguchi et al., 2007), the parameters of the Weibull distribution under the null are assumed to be: $k_0 = 1.2$ and $\lambda_0 = 16$. In the RMST, $\tau = 18$ is the time of interest.

The sample sizes are calculated for each test to attain 80% power at the significance level of 0.05 (one-sided). Statistical power is computed from simulations, and the sample size is the

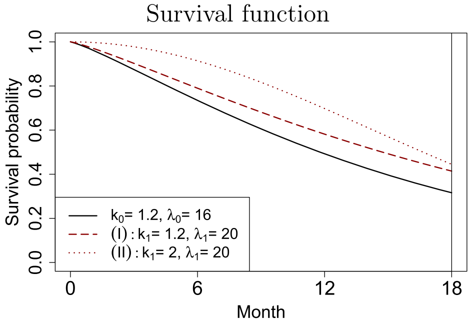

Figure 4: Survival curves for the example with two different alternative distributions.

Table 1: The computed sample sizes per treatment group in a balanced randomized trial to attain 80% power at the significance level of 0.05. The parameters of the Weibull distribution under the null are: $k_0 = 1.2$ and $\lambda_0 = 16$.

H_a distribution	RMST	logrank	Wilcoxon	CoxPH
(I) $k_1 = 1.2$ and $\lambda_1 = 20$	184	158	165	144
(II) $k_1 = 2.0$ and $\lambda_1 = 20$	57	88	59	81

smallest one whose simulated statistical power is above the nominal level. We consider two scenarios for the new treatment: (I) $k_1 = 1.2$ and $\lambda_1 = 20$; and (II) $k_1 = 2$ and $\lambda_1 = 20$. Figure 4 shows their survival curves. The first alternative distribution meets the PH assumption with $k_0 = k_1$, while the second alternative distribution does not and it has early benefit from the new treatment as compared to the existing treatment. The computed sample sizes per group are presented in Table 1. When the PH assumption is met, the CoxPH test needs the least sample size, followed by the logrank test, the Wilcoxon test, and the RMST test. The sample size saving is 22% by using the CoxPH test as compared to the RMST test. Under the second alternative distribution scenario with $k_1 = 2$, the PH assumption does not hold. The RMST test needs slightly fewer sample sizes than the Wilcoxon test. Their sample sizes are much smaller than those from the other two tests. As compared to the logrank test, the RMST test could save sample size by 35%. These sample size comparisons are consistent with the findings from simulation studies.

4 Discussion

The choice of the pre-specified time τ in the RMST would affect the statistical inference. The null curve can be estimated from historical data, but the survival curve of a new treatment is often unknown before the study. Even in some cases where a small number of patients treated by a new treatment could be used as pilot data, the variation of the survival curve still remains large. The chosen value of τ affects the difference in the RMST between a new treatment and the standard treatment (Shan et al., 2024; Shan, 2021; Lu et al., 2024). As pointed out by Tian et al. (2018), the value of τ should be less than the maximum possible observed times to provide valid statistical inference. The R function to compute simulated TIE and statistical power is included in the supplementary file.

A two-stage or multiple-stage design has the potential to save sample size when the patient accrual time t_a is relatively longer than the follow up time t_f to allow a certain percentage of participants completed the follow up before t_a (Shan and Zhang, 2019; Shan, 2020). When t_a is shorter than t_f, it is more likely that no patient has completed the final follow up before t_a. Then, all patients are already enrolled in the study. It is true that no sample size savings would occur, but the results from interim analysis could be useful in the future trials (Shan et al., 2016; Shan and Zhang, 2019; Jiang et al., 2020). Recently, Lu and Tian (2021) proposed a group sequential randomized clinical trial based on the RMST by using different τ values in computing the RMST at interim analysis. It is optimal to use the same τ at interim analysis as that in the final analysis, but it is a challenge with partially complete data at interim analysis.

Supplementary Material

The R function to compute simulated TIE and statistical power.

Acknowledgement

We would like to thank the comments from reviewers and associate editor, that help us to improve the manuscript.

Funding

Research reported in this publication was supported by the National Institutes of Health under Award Number R01AG070849 and R03AG083207.

References

Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 34(2): 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Harrington DP, Fleming TR (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3): 553–566. https://doi.org/10.1093/biomet/69.3.553

Jiang T, Cao B, Shan G (2020). Accurate confidence intervals for risk difference in meta-analysis with rare events. *BMC Medical Research Methodology*, 20(1): 98. https://doi.org/10.1186/s12874-020-00954-8

Liao JJ, Liu GF, Wu WC (2020). Dynamic RMST curves for survival analysis in clinical trials. *BMC Medical Research Methodology*, 20(1): 218. https://doi.org/10.1186/s12874-020-01098-5

Lu X, Zhang Y, Tang Y, Bernick C, Shan G (2025). Conversion to Alzheimer's disease dementia from normal cognition directly or with the intermediate mild cognitive impairment stage. *Alzheimer's & Dementia*, 21(1): e14393. https://doi.org/10.1002/alz.14393

Lu Y, Tian L (2021). Statistical considerations for sequential analysis of the restricted mean survival time for randomized clinical trials. *Statistics in Biopharmaceutical Research*, 13(2): 210–218. https://doi.org/10.1080/19466315.2020.1816491

Mantel N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports, Part 1*, 50(3): 163–170.

Peto R, Peto J (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A. General*, 135(2): 185. https://doi.org/10.2307/2344317

Royston P, Parmar MK (2013). 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 Medical Research Methodology*, 13: 152. https://doi.org/10.1186/1471-2288-13-152

Shan G (2020). Two-stage optimal designs based on exact variance for a single-arm trial with survival endpoints. *Journal of Biopharmaceutical Statistics*, 30(5): 797–805. https://doi.org/10.1080/10543406.2020.1730869

Shan G (2021). Optimal two-stage designs based on restricted mean survival time for a single-arm study. *Contemporary Clinical Trials Communications*, 21: 100732. https://doi.org/10.1016/j.conctc.2021.100732

Shan G (2022). Randomized two-stage optimal design for interval-censored data. *Journal of Biopharmaceutical Statistics*, 32(2): 298–307. https://doi.org/10.1080/10543406.2021.2009499

Shan G (2023). Response adaptive randomization design for a two-stage study with binary response. *Journal of Biopharmaceutical Statistics*, 33(5): 575–585.

Shan G, Dodge Francis C, Liu J, Hong X, Bernick C (2024). Application of adaptive designs in clinical research. In: *Modern Inference Based on Health-Related Markers: Biomarkers and Statistical Decision Making*, 229–243. Academic Press.

Shan G, Wilding GE, Hutson AD, Gerstenberger S (2016). Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine*, 35(8): 1257–1266.

Shan G, Zhang H (2019). Two-stage optimal designs with survival endpoint when the follow-up time is restricted. *BMC Medical Research Methodology*, 19: 74. https://doi.org/10.1186/s12874-019-0696-x

Shan G, Zhang Y, Tang Z, Ding A, Wu S (2025). Disease progression trajectory curves to estimate saved time in Alzheimer's disease trialsitle. *Contemporary Clinical Trials*, 151: 107814. https://doi.org/10.1016/j.cct.2025.107814

Takiguchi Y, Moriya T, Asaka-Amano Y, Kawashima T, Kurosu K, Tada Y, et al. (2007). Phase II study of weekly irinotecan and cisplatin for refractory or recurrent non-small cell lung cancer. *Lung Cancer*, 58(2): 253–259. https://doi.org/10.1016/j.lungcan.2007.06.004

Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, 74(2): 694–702. https://doi.org/10.1111/biom.12770

Uno H (2017). Vignette for survRM2 package: Comparing two survival curves using the restricted mean survival time. *Technical report*.

Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, et al. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Annals of Internal Medicine*, 163(2): 127–134. https://doi.org/10.7326/M14-1741

Zhang Y, Li Y, Song S, Li Z, Lu M, Shan G (2024). Predicting conversion time from mild cognitive impairment to dementia with interval-censored models. *Journal of Alzheimer's Disease*, 101(1): 147–157. https://doi.org/10.3233/JAD-240285