A nonparametric statistical method for two crossing survival curves

Xinghui Huang\textsuperscript{a}, Jingjing Lyu\textsuperscript{a}, Yawen Hou\textsuperscript{b}, and Zheng Chen\textsuperscript{a}

\textsuperscript{a}Department of Biostatistics, School of Public Health (Guangdong Provincial Key Laboratory of Tropical Disease Research), Southern Medical University, Guangzhou, China; \textsuperscript{b}Department of Statistics, College of Economics, Jinan University, Guangzhou, China

ABSTRACT

Background: In comparative research on time-to-event data for two groups, when two survival curves cross each other, it may be difficult to use the log-rank test and hazard ratio (HR) to properly assess the treatment benefit. Our aim was to identify a method for evaluating the treatment benefits for two groups in the above situation.

Methods: We quantified treatment benefits based on an intuitive measure called the area between two survival curves (ABS), which is a robust measure of treatment benefits in clinical trials regardless of whether the proportional hazards assumption is violated or two survival curves cross each other. Additionally, we propose a permutation test based on the ABS, and we evaluate the effectiveness and reliability of this test with simulated data.

Results: The ABS permutation test is a robust statistical inference method with an acceptable type I error rate and superior power to detect differences in treatment effects, especially when the proportional hazards assumption is violated.

Conclusion: The ABS can be used to intuitively quantify treatment differences over time and provide reliable conclusions in complicated situations, such as crossing survival curves.

Abbreviations: HR: hazard ratio; MT: median survival time; RMST: restricted mean survival time; ABS: area between two survival curves; \textsubscript{ABS}L/\textsubscript{ABS}R: area between two survival curves over the left/right time interval; \textsubscript{ABS}P: area between two survival curves permutation test; MTd: difference in median survival time; RMSTd: difference in restricted mean survival time

1. Background

Time-to-event end points are widely used in clinical follow-up studies. The difference in treatment benefits in survival analysis is commonly quantified and reported as the (cumulative) survival rate, the hazard ratio (HR), the median survival time (MT), or the restricted mean survival time (RMST) (Mantel \textit{et al.} 1966; Pepe and Fleming 1989; Royston and Parmar 2011; Valérie \textit{et al.} 2015; Rogawski \textit{et al.} 2016). The log-rank test corresponding to the HR is a common and classical choice when the HR between treatment...
arms remains constant over time (the proportional hazards assumption). However, when the proportional hazards assumption is violated, especially when two survival curves cross each other, the difference based on hazards may be offset before and after the crossing point, resulting in a loss of power of the log-rank test (Uno et al. 2014; Li et al. 2015). The method based on difference in the MT depends on information at the median survival point (when the survival rate is equal to 0.5) (Rogawski et al. 2016; Zhao et al. 2016). Therefore, when the censoring rate is large or the follow-up time is insufficient, this method cannot be applied without an accurate estimation of the MT (Saad et al. 2018). The method based on the difference in RMST does not require the assumption of proportional hazards (Dehbi, Royston, and Hackshaw 2017). However, when the two survival curves cross each other, the difference in the RMST may be offset before and after the crossing point, which causes an invalid assessment of the difference in treatment benefits and reduced power.

Hence, in this paper, we introduce a measure of the area between two survival curves (ABS) to interpret the difference in treatment benefits between two groups (Lin and Xu 2010; Seruga et al. 2012; Zhao et al., 2012). Moreover, we propose a corresponding test that can avoid the problems stated above (Stefania and Maria Grazia 2013). Simulations and examples are applied to compare the above measures. An R program has been developed to implement the ABS and the corresponding test.

2. Methods

The ABS is determined by the accumulated absolute difference in treatment effects from time 0 to the end of follow-up and can be estimated by integrating the absolute survival difference:

$$\text{ABS} = \int_0^v |S_1(t) - S_2(t)| dt \approx \sum_{j|t_j < v} |S_1(t_j) - S_2(t_j)| (t_{j+1} - t_j).$$

where $S_1(t), S_2(t)$ are the survival rates at time $t$ for each group, $t_j$ represents the pooled and ordered distinct event times in the two groups, and $v$ is the end of follow-up.

When comparing the differences between two groups, Lin and Xu (2010) estimated the mean $\bar{E}(\text{ABS})$ and variance $\bar{V}(\text{ABS})$ of the ABS under the null hypothesis and considered the normal transformation value $\Delta = \frac{\text{ABS} - \bar{E}(\text{ABS})}{\sqrt{\bar{V}(\text{ABS})}}$ to be a test statistic (ABS test) that would be an asymptotic standard normal distribution under the null hypothesis.

However, the test statistic $\Delta$ is non-normal in cases of small or large samples (details in Additional file 1). To provide a more robust confidence interval and a reliable hypothesis test result, we propose an ABS permutation test (ABSP test) based on permutation theory (Stefania and Maria Grazia 2013). The ABSP test generates its own critical regions and $P$ value directly from the permutation resample data rather than from the distribution of the test statistic under the null hypothesis. The ABSP test procedure is as follows. First, $\Delta$ is computed from the original data; second, the ABS and the test statistic $\Delta(1)$ are computed for the 1st iteration. All observations are combined from both groups, and the original data are randomly resampled without replacement. Third, step 2 is repeated an additional 999 times, and $\Delta(2), \Delta(3), \Delta(4), ..., \Delta(1000)$ is
Under the null hypothesis, with a mean of 0 and variance of \( \Delta(n) \) is satisfied and \( I(n) = 0 \) otherwise.

Since researchers may also be interested in the short-term or long-term treatment effects, we extended the ABS to any time interval (ABSi) and created the ABS permutation over time interval test (ABSi test). The null hypothesis is \( H_0 : S_2(t) = S_2(t), t^* < t < t'', t^* \geq 0, t'' \leq \tau \), where \( t^* \) and \( t'' \) are the left and right endpoints of the interval. The alternative hypothesis is \( H_1 : S_2(t) \neq S_2(t) \) for some \( t \) between \( t^* \) and \( t'' \). Then, the ABS over the time interval is given by

\[
\text{ABS}_{\text{int}} = \int_{t^*}^{t''} |S_1(t) - S_2(t)| \, dt \approx \sum_{j|t^* < t_j < t''} \left| S_1(t_j) - S_2(t_j) \right| (t_{j+1} - t_j).
\]

After the endpoints of the interval are converted to \([0, t'']\) via an indicator function \( \Phi(t_j > t^*) \), which is equal to 1 if \( t_j > t^* \) and equal to 0 otherwise, \( \text{ABS}_{\text{int}} \) can be expressed as

\[
\text{ABS}_{\text{int}} = \int_{t^*}^{t''} |S_1(t) - S_2(t)| \, dt = \int_0^{t''} \Phi(t > t^*) |S_1(t) - S_2(t)| \, dt \approx \sum_{j|t_j < t''} \left| S_1(t_j) - S_2(t_j) \right| \left[ \Phi(t_j > t^*) (t_{j+1} - t_j) \right].
\]

Therefore, based on the Kaplan-Meier estimation, we have

\[
\text{ABS}_{\text{int}} \approx \sum_{j|t_j < t''} \left| \hat{S}_1(t_j) - \hat{S}_2(t_j) \right| \left[ \Phi(t_j > t^*) (t_{j+1} - t_j) \right].
\]

Under the null hypothesis, \( \hat{S}_1(t) - \hat{S}_2(t) \) follows an approximately normal distribution with a mean of 0 and variance of \( \left[ \hat{\sigma}_S^2(t) + \hat{\sigma}_{S_2}^2(t) \right] \). Then, we have the following:

\[
\hat{E}\left( \left| \hat{S}_1(t) - \hat{S}_2(t) \right| \right) = \left\{ \frac{2}{\pi} \left[ \hat{\sigma}_S^2(t) + \hat{\sigma}_{S_2}^2(t) \right] \right\}^{1/2},
\]

\[
\hat{V}\left( \left| \hat{S}_1(t) - \hat{S}_2(t) \right| \right) = \left( 1 - \frac{2}{\pi} \right) \left[ \hat{\sigma}_S^2(t) + \hat{\sigma}_{S_2}^2(t) \right].
\]

Based on the normal approximation, we can estimate the mean and variance of \( \text{ABS}_{\text{int}} \) as follows:

\[
\hat{E}(\text{ABS}_{\text{int}}) = \sum_{j|t_j < t''} \left\{ \frac{2}{\pi} \left[ \hat{\sigma}_S^2(t_j) + \hat{\sigma}_{S_2}^2(t_j) \right] \right\}^{1/2} \Phi(t_j > t^*) (t_{j+1} - t_j),
\]

\[
\hat{V}(\text{ABS}_{\text{int}}) = \sum_{j|t_j < t''} \Phi(t_j > t^*) (t_{j+1} - t_j)^2 \left( 1 - \frac{2}{\pi} \right) \left[ \hat{\sigma}_S^2(t_j) + \hat{\sigma}_{S_2}^2(t_j) \right]
\]

\[
+ \sum_{j < j'|t_j, t_j' < t''} 2 \rho_{j,j'} \Phi(t_j > t^*) \Phi(t_{j'} > t^*) (t_{j+1} - t_j) (t_{j'+1} - t_{j'}) \left( 1 - \frac{2}{\pi} \right)
\]

\[
\times \left\{ \left[ \hat{\sigma}_S^2(t_j) + \hat{\sigma}_{S_2}^2(t_j) \right] \left[ \hat{\sigma}_S^2(t_{j'}) + \hat{\sigma}_{S_2}^2(t_{j'}) \right] \right\}^{1/2}.
\]
where \( q_{j,j_0} \) is the correlation coefficient between \( \hat{S}_1(t_{j_0})/C_0 \) and \( \hat{S}_2(t_{j_0})/C_1/C_2/C_2/C_2/C_2 \) and \( j < j_0 \). Lin and Xu (2010) suggested setting \( q_{j,j_0} \) to 0.5 for all \( j \) and \( j_0 \). Therefore, we have the test statistic \( \Delta_{int} = [ABS_{int} - \hat{E}(ABS_{int})]/\sqrt{\hat{V}(ABS_{int})} \). Using permutation theory, we can compute the \( P \) value based on 1000 resamples.

3. Simulation study

3.1. Simulation design

To evaluate and compare the statistical performance of the ABSP test with that of the log-rank test, the RMST test, and the ABS test, we calculated the type I error rates and power via Monte Carlo simulations (the test based on the difference in MT is excluded here because the survival rate of any group may not reach 0.5, which cannot yield a test result). We considered the following situations (shown in Figure 1): Situation I. two survival curves overlapping (the differences between two groups are generated by random sampling to evaluate false positive rates of tests under the null hypothesis, that is, type I error rates); Situation II. two survival curves with proportional hazard functions; Situation III. two survival curves with early and middle differences; Situation IV. two survival curves with early and middle differences. Specifically, Situation I. \( \lambda_1 = \lambda_2 = 0.25 \). Situation II. \( \lambda_1 = 0.5 \) and \( \lambda_2 = 0.2 \). Situation III. \( \lambda_1 = 1.2, 0.1, 0.5, 1 \) and \( \lambda_2 = 0.5, 0.1, 1.2, 1 \) for \( t \leq 0.8, 0.8 < t \leq 1.8, 1.8 < t \leq 2.6, t > 2.6 \), respectively. Situation IV. \( \lambda_1 = 1, 2.5 \) and \( \lambda_2 = 1, 0.5 \) for \( t \leq 0.8, t > 0.8 \), respectively. Situation V. \( \lambda_1 = 1/12 \) and \( \lambda_2 = 1/4, 1/35 \) for \( t \leq 2, t > 2 \), respectively. Situation VI. \( W(1.5, 5) \) for group 1; \( \lambda = 0.5, 0.1 \) for \( t \leq 1.5, t > 1.5 \), respectively, for group 2.
survival curves with a late difference; Situation V. two survival curves with median crossing; and Situation VI. two survival curves with late crossing.

We considered censoring rates of 0%, 15%, 30%, and 45% as well as groups of equal sample sizes ($N_1 = N_2 = 20, 50, 100$) or unequal sample sizes ($N_1 = 20, N_2 = 50; N_1 = 50, N_2 = 100$). The lifetime $X$ was generated from an exponential or Weibull distribution, and the variable $status$ was set as $\delta = 1$ when there was no censoring. When we considered right-censored data, we assumed that patients entered the study uniformly over a time interval and that censoring time $C$ was generated from uniform distributions $U(0, a)$ and $U(0, b)$. We denoted the lifetime $T = \min(X, C), \delta = I[X \leq C]$, where $I(\cdot)$ is an indicator function. We could adjust the average censoring rates through the values of $a$ and $b$. The number of iterations in each simulation was 1000, and we would reject the null hypothesis at the 0.05 significance level.

### 3.2. Simulation results

To present the simulation results clearly, we provide the summarized results of the type I error rates (Table 1) and power for several tests (Table 2), where the summarized results are obtained by variance (ANOVA) techniques (original results in Additional file 2) (Klein et al. 2007; Logan, Klein, and Zhang 2008). To evaluate the type I error rates, the response variable $Y$ was defined as the percent rejection rate minus the nominal 5% level; therefore, good performance is implied by close-to-zero estimates for the expectation $E(Y)$ in the ANOVA. To evaluate power, the outcome variable $Y$ was defined as the rejection rate: good performance is indicated by large estimated values of $E(Y)$. Here, we consider four factors: $TEST$, with 4 levels; $NUM$, with 5 levels; $CENSORE$, with 4 levels; and $SITUATION$, with 6 levels. These factors represent the test method, sample size of each group, censoring rate, and simulation situation, respectively. We fit four models without intercepts for $E(Y)$:

- **Model 1:** $E(Y) = TEST \times NUM + CENSORE + SITUATION$.
- **Model 2:** $E(Y) = TEST \times CENSORE + NUM + SITUATION$.
- **Model 3:** $E(Y) = TEST \times SITUATION + CENSORE + NUM$.
- **Model 4:** $E(Y) = TEST + SITUATION + CENSORE + NUM$.

| NUM | TEST | Log-rank | RMSTd | ABS | ABSP |
|-----|------|----------|-------|-----|------|
| (20, 20) | -0.650 | -1.450 | -2.300 | -1.250 |
| (50, 50) | -0.425 | -0.050 | -1.825 | 0.050 |
| (100, 100) | 0.250 | -0.300 | -2.500 | -0.250 |
| (20, 50) | -0.125 | 0.000 | -1.000 | -0.250 |
| (50, 100) | -0.625 | -0.475 | -2.150 | -0.575 |
| CENSORE | ¶ | 0.000 | 0.020 | 0.960 | 2.820 | 0.540 |
| | 0.15 | -0.020 | -0.940 | -2.580 | -0.500 |
| | 0.30 | -0.480 | 0.040 | -1.520 | -0.280 |
| | 0.45 | -0.060 | 0.040 | -0.900 | -0.500 |
| SITUATION | ¶ | 0.145 | -0.455 | -1.955 | -0.455 |
| TOTAL§ | -0.145 | -0.455 | -1.955 | -0.455 |

*: Model 1; ¶: Model 2; §: Model 3; ¶: Model 4; CENSORE: censoring rate; RMSTd: test based on the difference in the RMST; ABS: ABS test; ABSP: ABS permutation test; default range of average deviations for 1000 iterations: [-1.351, 1.351].
As shown in Table 1, the type I error rates of the ABS test deviate from the prespecified nominal level, whereas those of the other tests are close to 0.05. The censoring rates and sample sizes have little impact on the type I error rate.

As shown in Table 2, in situation II, the four tests have similar mean rejection rates, and the log-rank test performs best. The ABSP test is robust and has the highest mean rejection rates in other situations. Whether the sample sizes of the two groups are equal has little effect on the power.

The log-rank test may have worse power when the proportional hazards assumption is violated. Tests based on a difference in RMST may lose power when the survival curves cross because of differences in offsetting before and after the crossing point. The ABS test is not reliable because the type I error rates deviate from the prespecified nominal level. The ABSP test is a robust test for comparing treatment benefits, regardless of whether the proportional hazards assumption is violated or the survival curves cross.

### 3.3. Examples

**Example 1:** In a clinical trial on the role of docetaxel in non-metastatic castration-resistant prostate cancer (nmCRPC) (Ito et al. 2018), 46 patients received docetaxel for nmCRPC, and 52 had distant metastases. As shown in Figure 2A (simulated data from Figure 3A in the original paper), the Kaplan-Meier curves of overall survival in the two groups are separate, and the proportional hazards assumption is satisfied \((P = 0.881)\). As shown in Example 1 in Table 3, the test based on the difference in the MT cannot provide a result because the survival rate of the nmCRPC group did not reach 0.5, whereas the other tests indicate that the treatment effect in the nmCRPC group was better than that in the placebo group. Thus, the treatment effect in the nmCRPC group was better than that in the mCRPC group in both the short term and the long term (Example 1 in Table 3).

**Example 2:** The kidney-dialysis trial was designed to assess the time to the first exit-site infection (in months) in patients with renal insufficiency (Yan 2004). In 43 patients, the catheter was surgically implanted, whereas in 76 patients, the catheter was percutaneously placed. There were 27 patients censored in the first group and 65 in the second.
As shown in Figure 2B, the two survival curves cross at an early time, and the Grambsch-Therneau test indicates that the proportional hazards assumption is violated \((P = 0.003)\). As shown in Example 2 in Table 3, the test based on the difference in the MT cannot provide a result as in Example 1, and the log-rank test indicates no significant difference, whereas the test based on the difference in the RMST and the ABSP test yield significant overall results. Further, the treatment effect in the percutaneously placed catheter group is better than that in the surgically implanted catheter group in the long term (after 8 months), which may result in an overall significant difference.

Example 3: In a randomized clinical trial on arginine deprivation with pegylated arginine deiminase in patients with argininosuccinate synthetase 1-deficient malignant pleural mesothelioma (Szlosarek et al. 2017), 68 patients were randomized to receive the arginine-lowering agent pegylated arginine deiminase (ADI-PEG20, 36.8 mg/m², weekly, intramuscular) plus best supportive care (Group 1) or best supportive care alone (Group 2). As shown in Figure 2C (simulated data from Figure 2B in the original paper with approximately the same sample size, censoring rate, and basic statistics; the sample sizes of Group 1 and Group 2 were 44 and 24 patients, and the censoring rates in the two groups were approximately 4.5% and 4.2%, respectively), the two overall survival curves cross in the 10th month, and the survival rates of the two groups reverse before and after the crossing point, which may be caused by the treatment effects between the
two groups differing from the early to late phases. The Grambsch-Therneau test indicates that the proportional hazards assumption is violated ($P = 0.029$).

The descriptive statistics and test results based on the HR, MT, RMST, and ABS are shown in Table 3. The HR of Group 1 is $0.673$ times that of Group 2 (HR $= 0.673$; 95% CI: 0.394 to 1.148; $P = 0.147$); the difference in the MT between groups is 0.120 (95% CI: $-0.064$ to 0.308; $P > 0.999$) months; and the difference in RMST between the two groups is 3.714 (95% CI: $-1.126$ to 8.553; $P = 0.133$) months. These three measures indicate no statistical significance for the overall test. However, the ABS, which reflects the absolute survival difference between the two groups, is 5.440 (95% CI: 2.914 to 8.888) months, indicating a significant difference (ABSP $P = 0.031$). Since the two survival curves cross in the 10th month, which indicates that the treatment effects of the two groups may reverse before and after the crossing point, we used the ABSPi test to analyze the interval differences before and after the crossing point. As shown in Table 3, the ABSPi test indicates that the short-term treatment effects of the two groups are not significantly different before the 10th month from the start of follow-up ($P = 0.360$), whereas the differences in long-term treatment effects of the two groups are statistically significant. The treatment effect of Group 1 is better than that of Group 2 at the 10th month or later ($P = 0.028$). In other words, no difference between treatments is observed in the short term, while the treatment effect in Group 1 is better than that in Group 2 in the long term, resulting in an overall significant difference (ABSP $P = 0.031$).

4. Discussion

When the proportional hazards assumption is violated, especially when two survival curves cross, the HR may change over time; thus, a constant HR cannot be used to interpret the differences in overall survival. Therefore, the corresponding log-rank test is not appropriate. A test based on a difference in MT, which does not require the proportional hazards assumption and uses information only at the MT, may lose information and may be affected by the scheduled endpoint assessment rather than overall survival (Uno et al. 2014; Rogawski et al. 2016; Saad et al. 2018). Meanwhile, a test based on a difference in MT is not applicable when the survival rate does not reach 0.5 or when the crossing point of the survival curves is located near 0.5. A test based on the difference in the RMST is easily calculated and interpreted according to the difference in areas under survival curves (Royston and Parmar 2011). This test is typically not affected by the proportional hazards assumption; however, when the survival curves cross, as in Example 2 and Example 3 (Figures 2B/2C), the RMSTs of Group 1 and Group 2 can be expressed as “area $D_2+S$” and “area $D_1+S$”. Thus, the difference in the RMST can be estimated as “$D_2-D_1”$, which is the same as the differences in the ABS before and after the crossing point. Therefore, a test based on the difference in the RMST gives only the relative difference in effects and does not lose power in this case because of differences in offsetting before and after the crossing point (the same occurs in Situations V and VI). However, the ABS can be expressed as “area $D_1+D_2$”, which reflects the absolute difference in treatment effects from the start to the end of follow-up. The ABS is not affected by violation of the proportional hazards assumption or
crossing of the survival curves. When the survival curves are separated, a test based on the ABS yields a similar result to the test based on a difference in the RMST.

5. Conclusions

The ABS measure effectively reflects the absolute benefit of treatment effects between two groups in randomized clinical trials. The ABS is a robust measure for comparing the treatment benefits of two groups and can be applied regardless of whether the proportional hazards assumption is violated or the survival curves cross. Meanwhile, the corresponding ABSP test is also robust and reliable. When researchers are interested in treatment effects over any time interval (i.e., evaluation of the short-term or long-term treatment benefits of two cancer drugs), the corresponding ABSPi test can be applied in combination with Kaplan-Meier curves to draw a comprehensive conclusion. Additionally, an R program was developed for the application of the ABS and the corresponding test (details in Additional file 3).

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by the National Natural Science Foundation of China (81673268, 81903411) and Natural Science Foundation of Guangdong Province (2017A030313812, 2018A030313849).

References

Dehbi, H.-M., P. Royston, and A. Hackshaw. 2017. Life expectancy difference and life expectancy ratio: Two measures of treatment effects in randomised trials with non-proportional hazards. BMJ (Clinical Research ed.) 357 (357):j2250 doi:10.1136/bmj.j2250.

Ito, K., T. Kimura, H. Onuma, R. Tabata, T. Shimomura, K. Miki, M. Tomita, and S. Egawa. 2018. Does docetaxel prolong survival of patients with non-metastatic castration-resistant prostate cancer? The Prostate 78 (7):498–505. doi:10.1002/pros.23493.

Klein, J. P., B. Logan, M. Harhoff, and P. K. Andersen. 2007. Analyzing survival curves at a fixed point in time. Statistics in Medicine 26 (24):4505–19. doi:10.1002/sim.2864.

Li, H., D. Han, Y. Hou, H. Chen, and Z. Chen. 2015. Statistical inference methods for two crossing survival curves: A comparison of methods. PloS One 10 (1):e0116774. doi:10.1371/journal.pone.0116774.

Lin, X., and Q. Xu. 2010. A new method for the comparison of survival distributions. Pharmaceutical Statistics 9 (1):67–76. doi:10.1002/pst.376.

Logan, B. R., J. P. Klein, and M.-J. Zhang. 2008. Comparing treatments in the presence of crossing survival curves: An application to bone marrow transplantation. Biometrics 64 (3):733–40. doi:10.1111/j.1541-0420.2007.00975.x.

Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50 (3):163–70.

Pepe, M. S., and T. R. Fleming. 1989. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. Biometrics 45 (2):497–507. doi:10.2307/2531492.
Rogawski, E. T., D. J. Westreich, G. Kang, H. D. Ward, and S. R. Cole. 2016. Estimating differences and ratios in median times to event. *Epidemiology* 27 (6):848–51.

Royston, P., and M. K. Parmar. 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 30 (19):2409–21. doi:10.1002/sim.4274.

Saad, E. D., J. R. Zalcberg, J. Péron, E. Coart, T. Burzykowski, and M. Buyse. 2018. Understanding and communicating measures of treatment effect on survival: Can we do better? *JNCI: Journal of the National Cancer Institute* 110 (3):232–40. doi:10.1093/jnci/djx179.

Seruga, B., G. R. Pond, P. C. Hertz, E. Amir, A. Ocana, and I. F. Tannock. 2012. Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. *Annals of Oncology* 23 (11):2977–82. doi:10.1093/annonc/mds174.

Stefania, G., and V. Maria Grazia. 2013. Multivariate permutation test to compare survival curves for matched data. *BMC Medical Research Methodology* 13 (1):16.

Szlosarek, P. W., J. P. Steele, L. Nolan, D. Gilligan, P. Taylor, J. Spicer, M. Lind, S. Mitra, J. Shamash, M. M. Phillips, et al. 2017. Arginine deprivation with pegylated arginine deiminase in patients with argininosuccinate synthetase 1-deficient malignant pleural mesothelioma: A randomized clinical trial. *JAMA Oncology* 3 (1):58. doi:10.1001/jamaoncol.2016.3049.

Uno, H., B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, et al. 2014. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* 32 (22):2380–5. doi:10.1200/JCO.2014.55.2208.

Valérie, G., A. Sandrine, et al. 2015. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Statistics in Medicine* 34 (4):541–57.

Yan, R. B. J. 2004. Survival analysis: Techniques for censored and truncated Databy John P. Klein; Melvin L. Moeschberger. *Journal of the American Statistical Association* 99 (467):900–1. doi:10.1198/jasa.2004.s348.

Zhao, L., B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. J. Wei. 2016. On the restricted mean survival time curve in survival analysis. *Biometrics* 72 (1):215–21. doi:10.1111/biom.12384.

Zhao, L., L. Tian, H. Uno, S. D. Solomon, M. A. Pfeffer, J. S. Schindler, and L. J. Wei. 2012. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring and analyzing a comparative clinical study. *Clinical Trials: Journal of the Society for Clinical Trials* 9 (5):570–7. doi:10.1177/1740774512455464.