The Welch-Satterthwaite Adjustment for Censored Data Two Sample Restricted Mean Tests

Mai Zhou*

Abstract

The Welch-Satterthwaite approximate two sample t-test is generally preferred over the pooled variance version, because equal variances assumption is seldom true. We propose to use the Welch-Satterthwaite adjustment for two-sample censored data Restricted Mean Survival Time tests. We apply the Welch-Satterthwaite adjustment to both the Wald tests and Wilks tests. Simulations are provided which show the adjustment significantly improve the small sample accuracy of the tests.

MSC 2010 Subject Classification: Primary 62N02; secondary 62G05, 62B10.

Key Words and Phrases: Restricted mean survival times, Small samples, Accuracy of type I error.

1 Introduction

The log-rank test is often used for compare two samples of right censored. However, its performance may seriously degrade when two hazards are not proportional to each other. When the proportional hazards assumption is in doubt, many researchers recently propose to use the restricted mean survival time (RMST) as a measure to evaluate/compare treatments with right censored data. See for example [1], [2], [3], [4], [5] to name a few.

Some simulation show that the tests based on the RMST can be very competitive to the log-rank tests when the proportional hazards assumption does hold, but much better than the log-rank test when the proportional hazards assumption is violated, [16]. Also the restricted mean value is an intuitive and easy to interpret characteristic/measure.

However, simulations also show that the Wald type RMST tests have poor accuracy (inflated type I error) for small sample sizes [15]. Similar problem also exist (though less profound) if we use the Wilks type tests based on empirical likelihood [17], [18].

*Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, KY 40536 USA

In order to improve small sample accuracy, we propose to replace the standard normal distribution in the Wald tests by a t-distribution with the Welch-Satterthwaite degrees of freedom [7], [8]. Similarly, for the Wilks tests based on empirical likelihood, we replace the standard chi square distribution by a t-distribution square, with the Welch-Satterthwaite DF.

Owen (2001), in the context of one-sample tests, suggest to fine tuning the smaller sample accuracy of the empirical likelihood test, by use a t-distribution square (or F-distribution) to replace the typical chi-square limiting null distribution for the empirical likelihood ratio test. This is a level two adjustment as opposed to the W-S, which is a level three adjustment (see section 2).

Simulations are conducted which show the W-S adjustment improves the accuracy of error rate for both the Wald and Wilks two sample tests, especially for smaller sample sizes.

2 Two Sample Welch-Satterthwaite Adjustment

We can identify three levels of accuracy when trying to come up with an approximate null distribution for a test statistic.

Level one is to use a single (limiting) distribution, no matter what sample sizes are and no matter how the sample data at hand look like. An example is the approximate 95% (Wald) confidence interval (in no censor data case)

$$\bar{X} - \bar{Y} \pm 1.96\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} \ .$$

Here the coefficient 1.96 stems from a **single** standard normal distribution.

Level two would incorporate the sample size into the construction of the distribution (t-distribution or F-distribution with DF related to current sample size). An example is the approximate 95% confidence interval

$$\bar{X} - \bar{Y} \pm t_{n+m-2}(0.975)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} \ .$$

Here $t_{n+m-2}(0.975)$ incorporate the information of sample sizes n and m. In other words, the distribution in question is not a single one but a series of t-distributions with different DFs depend on the sample sizes.

A further improvement, level three, would also use the sample data in addition to the sample size in constructing the distribution. An example is the approximate 95% confidence interval

$$\bar{X} - \bar{Y} \pm t_{\nu}(0.975)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} \ ,$$

where the degree of freedom ν is the so called Welch-Satterthwaite DF:

$$\nu = \frac{\left(s_x^2/n + s_y^2/m\right)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)} \ . \tag{1}$$

Notice the calculation of ν uses not only the sample sizes n and m but also the sample data in the form of s_x^2 and s_y^2. Thus ν is random (as opposed to the distributions in Level one and two). Another example of Level three approximation is bootstrap, which also uses sample observation to construct a (random) distribution, conditional on the given sample. We shall not delve into this topic here.

We point out that the (empirical) likelihood ratio based tests/confidence intervals (Wilks confidence intervals), are also examples of Level 3 approximation by above categorization. Since the shape of the Wilks confidence interval/region is random (depend on the given sample data) even for fixed sample size.

The above examples are for non-censored data. But the same idea apply for censored data: where we try to test/estimate the difference of two restricted mean survival times. The null distribution of the test statistic can be better approximated by techniques of level 3, if done right.

Another related phenomena is the expected verses observed Fisher information: observed information is a level 3 approximation while the expected information is level one. Efron and Hinkeley (1978) have argued that using the observed information actually leads to better approximations.

3 Analysis of two-sample RMSTs

In the analysis of RMST, the exact distribution of the integrated Kaplan-Meier is unknown, and we use the fact that the distributions are asymptotically normal. Further, the assumption of equal variances for two samples are not reasonable. Therefore a Welch-Satterthwaite calibrations is appropriate.

3.1 Two-sample Wald Confidence Interval for RMST

The exact formula for the Wald test is detailed in [15], including the variance estimator of the integrated Kaplan-Meier estimator. The general reference paper here is Uno et. al. (2014) and calculation is available in R package `survRM2`.

Our proposed adjustment simply replace the standard normal distribution with the t-distribution, degrees of freedom ν as in (1).

3.2 Two-sample Wilks Confidence Interval for RMST

The reference paper here is Zhou (2020) [17]. For the empirical likelihood tests in general, references are Owen 2001 [6] and Zhou 2016 [18]

Here the construction of confidence interval (or test) do not need the estimated variance. We use the same W-S degree of freedom calculated from previous subsection (Wald test) here to identify the t-distribution for our significance calculation. In other words, the -2 log Empirical Likelihood Ratio is deemed significant at 5% level, if and only if it is larger than $[qt(0.975, df = \nu)]^2$, etc. Here $qt(0.975, n)$ denote the 0.975 quantile of a t-distribution with degrees of freedom n.

4 Simulations

We focus on the testing of two sample Restricted Mean Survival Time difference with right censored data.

In Table 1, data were generated under null hypothesis of equal RMSTs. Actual type I error at nominal 5% and 1% errors are shown. With different sample sizes. Based on 5000 simulation runs.

Sample 1: Weibull shape=0.9, scale =12.7

Sample 2: Weibull shape =0.9, scale =12.7

Censoring distribution for both sample are Uniform (0, 12.5).

sample size	restriction	No calib.	W-S calib.	No calib.	W-S calib.
30, 20	$\tau = 10$	6.87%	6.13%	2.06%	1.62%
35, 25	$\tau = 10$	6.51%	5.96%	1.77%	1.47%

Table 1. Welch-Satterthwaite calibration for Wald test from `survRM2`. Nominal 5% and 1%. Under H_0.

Next, same simulation but under the alternative hypothesis:

Sample 1: Weibull shape=1.9, scale =12.7

Sample 2: Weibull shape =0.9, scale =12.7

Same censoring distributions as above.

Table 2 contains the actual observed error, i.e. percentage of confidence intervals NOT covering the true RMST difference (here true RMST differences are not zero): for nominal 95% confidence intervals and 99% intervals.

sample size	restriction	No calib.	W-S calib.	No calib.	W-S calib.
25, 20	$\tau = 10$	7.47%	6.56%	2.19%	1.58%
30, 25	$\tau = 10$	6.86%	6.23%	1.71%	1.36%
30, 20	$\tau = 10$	7.29%	6.5%	2.31%	1.76%
35, 25	$\tau = 10$	6.83%	6.19%	1.95%	1.56%

Table 2. Welch-Satterthwaite correction for Wald test from `survRM2`. Under H_A.

Next, we apply the W-S correction to the empirical likelihood test/confidence interval (Wilks interval) with the same data as above.

Table 3 is the results under H_0. Different sample sizes. Chi-square entry is for no adjustment (Level 1), $t^2(n1 + n2 - 2)$ entry is a level 2 adjustment and W-S entry is a level 3 adjustment discussed above.

Nominal	sample size	restriction	Chi-square	$t^2(n1 + n2 - 2)$	W-S calib.
5%	35, 25	$\tau = 10$	6.54%	6.04%	5.94%
1%	35, 25	$\tau = 10$	1.64%	1.34%	1.28%

Table 3. Simulated type I error of EL test using Chi-square, $t^2(n1 + n2 - 2)$ and Welch-Satterthwaite calibration

Table 4 is the results for Empirical likelihood ratio RMSTs tests, but under the same H_A as in table 2. (based on 5000 runs)

Nominal	sample size	restriction	Chi-square	$t^2(n1 + n2 - 2)$	W-S calib.
5%	35, 25	$\tau = 10$	6.7%	6.12%	5.9%
1%	35, 25	$\tau = 10$	1.76%	1.4%	1.32%

Table 4. Simulated error of EL test using Chi-square, $t^2(n1 + n2 - 2)$ and Welch-Satterthwaite

We can see that

(1) the W-S calibration always improves the accuracy of the approximation of actual error to nominal error, for either the Wald test or the Wilks test.

(2) The W-S calibration improvements are more profound under alternative hypothesis and for unequal sample sizes.

(3) Even with the W-S calibration, the actual error of Wilks tests are still larger than the nominal values (more so for Wald tests).

(4) With W-S calibration the Wilks tests are often more accurate than the Wald tests. (e.g. Under H_A 5.9% vs. 6.19% after W-S calibration. Under H_0 5.94% vs. 5.96% after W-S calibration).

5 Discussion

The Welch-Satterthwaite calibration is simple enough that it almost cost nothing to compute. The improvement for error is robust and across board so we recommend it for *all* two sample RMST tests.

In counting the sample size for either sample 1 or sample 2, i.e. $N1$ or $N2$, there are some rare cases that a right censored observation occurs before any real failure. We recommend delete this censored observation instead of counting it in the sample size. The reason is that such observation does not carry any information in the nonparametric inference such as the Kaplan-Meier.

Bartlett correction [6] is another possibility for improving small sample accuracy of Wilks tests but it seems too complicated to work reliably in censored data setting, if at all.

References

[1] Royston P, Parmar MKB. (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*. 30:2409–2421.

[2] Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, Wei LJ. (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*. 9:570–577.

[3] Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. and Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32, pp. 2380–2385.

[4] Kim DH, Uno H, Wei LJ. (2017). Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol*; 2(11): 1179–1180. doi: 10.1001/jamacardio.2017.2922

[5] Abulizi, X; Ribaudo, H J.; Flandre, P. (2019). The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trial: Reanalysis of the ACTG A5257 Trial. *Journal of Acquired Immune Deficiency Syndromes* Vol. 81 Issue 1 pp. 44–51. doi: 10.1097/QAI.0000000000001978

[6] Owen, A. (2001). *Empirical Likelihood* Chapman & Hall/CRC Press.

[7] Satterthwaite, F.E. (1946), An Approximate Distribution of Variance Components, *Biometrics Bulletin*, 2: 110-114.

[8] Welch, B.L. (1947), The Generalization of Students's Problem when Several Different Population Variances are Involved, *Biometrika*, 34: 28-35.

[9] Benedetti, J., Liu, P-Y., Sather, H., Seinfeld, J. and Epton, M. (1982). Effective Sample Size for Tests of Censored Survival Data. *Biometrika* Vol. 69, No. 2, 343–349.

[10] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of tha American Statistical Association* 53, 457–481.

[11] Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher Information. *Biometrika*. 65 (3): 457–487.

[12] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap* Chapman & Hall/CRC Press.

[13] Klein, JP. (1991). Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. *Scandinavian Journal of Statistics*, 18, 333–340.

[14] Meeker, WQ and Escobar, L. (1995). Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation *The American Statistician*, Vol. 49, No. 1, pp. 48–53.

[15] Horiguchi, M. and Ono, H. (2020). On permutation tests for comparing restricted mean survival time with small sample from randomized trials. *Statistics in Medicine* doi:10.1002/sim.8565.

[16] Zhou, M. (1992). Difference of means test with censored Data. *Communications in Statistics* 21, 697-706.

[17] Zhou, M. (2020). Restricted mean survival time and confidence intervals by empirical likelihood ratio. *Journal of Biopharmaceutical Statistics* https://doi.org/10.1080/10543406.2020.1862143

[18] Zhou, M. (2016). *Empirical Likelihood Methods in Survival Analysis* CRC Press, Taylor & Francis Group, New York.

6 R functions that compute the Welch-Satterthwaite effective degrees of freedom.

First, we need the function `rmst2_update.WS` which is a modification of an existing one in the R package `survRM2`, by adding the calculation and output of a W-S effective DF adjusted P-value.

Second, we need the function `rmst1_update` (from the R package `survRM2`), which is used by `rmst2_update.WS`. The R package `survival` is also needed.

```
rmst2_update.WS <- function(time, status, arm, tau=NULL, alpha=0.05,
                  var.method="greenwood", test, Mean0=0){

Z=list()

#--
wk1=rmst1_update(time[arm==1], status[arm==1], tau, alpha, var.method)
wk0=rmst1_update(time[arm==0], status[arm==0], tau, alpha, var.method)

#--- contrast (RMST difference) ---
rmst.diff.10     = wk1$rmst[1]-wk0$rmst[1] - Mean0   #### Added by Mai Zhou 8/2020
rmst.diff.10.se  = sqrt(wk1$rmst.var + wk0$rmst.var)
rmst.diff.z      = rmst.diff.10/rmst.diff.10.se

######--- Added by Mai Zhou, 8/2020 ---
N0 <- length(time[arm==0])
N1 <- length(time[arm==1])
s0dN0 <- wk0$rmst.var
s1dN1 <- wk1$rmst.var
effDF <- (s0dN0 + s1dN1)^2/((s0dN0)^2/(N0-1) + (s1dN1)^2/(N1-1))
WS.Pval.2side <- 2*pt(-abs(rmst.diff.z), df=effDF)
Calib.out = c(effDF, N0-1, N1-1, WS.Pval.2side)
names(Calib.out) = c("effDF", "DF0", "DF1", "Welch-Satterthwaite.pval.2side")
######-------------------------

if(test=="1_side"){
rmst.diff.z.1side      = rmst.diff.z
rmst.diff.pval.1side   = 1-pnorm(rmst.diff.z.1side) # one-sided test (upper)
rmst.diff.result.1side = cbind(rmst.diff.10, rmst.diff.10.se, 1,
              rmst.diff.z.1side, rmst.diff.pval.1side, wk1$rmst[1], wk0$rmst[1])
#--
out = rmst.diff.result.1side
}else{
```

```
#test=="2_side"
rmst.diff.z.2side      = abs(rmst.diff.z)
rmst.diff.pval.2side   = pnorm(-rmst.diff.z.2side)*2 # two-sided test
rmst.diff.result.2side = cbind(rmst.diff.10, rmst.diff.10.se, 2,
              rmst.diff.z.2side, rmst.diff.pval.2side, wk1$rmst[1], wk0$rmst[1])
#--
out = rmst.diff.result.2side
}

#--- results ---
rownames(out)=c("RMST (arm=1)-(arm=0)")
colnames(out) = c("Est.", "S.E.", "test-side", "z", "p", "rmst1", "rmst0")

#--- output ---
Z$unadjusted.result = out
Z$T.dist.calib = Calib.out
class(Z)="rmst2_update"

Z
}
```

We added an extra input `Mean0` in the function `rmst2_update.WS` so that the function can be used to test an alternative hypothesis. Default value `Mean0=0` which is the null hypothesis.

```
rmst1_update <- function(time, status, tau, alpha=0.05, var.method="greenwood"){

ft  = survfit(Surv(time, status)~1)   #
idx = ft$time<=tau

wk.time = sort(c(ft$time[idx],tau))
wk.surv = ft$surv[idx]
wk.n.risk  = ft$n.risk[idx]
wk.n.event = ft$n.event[idx]

time.diff = diff(c(0, wk.time))
areas = time.diff * c(1, wk.surv)
rmst = sum(areas)
rmst

#--- asymptotic variance ---
if(var.method=="greenwood"){
#--Greenwood plug-in estimator
```

```
wk.var <- ifelse((wk.n.risk-wk.n.event)==0, 0,
                  wk.n.event /(wk.n.risk *(wk.n.risk - wk.n.event)))
}
if(var.method=="aj"){
#--Aalen-Johansen plug-in estimators
wk.var <- ifelse( wk.n.risk==0, 0, wk.n.event /(wk.n.risk *wk.n.risk))
}

wk.var = c(wk.var,0)
rmst.var = sum( cumsum(rev(areas[-1]))^2 * rev(wk.var)[-1])
rmst.se  = sqrt(rmst.var)

#--- output ---
out=matrix(0,2,4)
out[1,]=c(rmst, rmst.se, rmst-qnorm(1-alpha/2)*rmst.se, rmst+qnorm(1-alpha/2)*rmst.se)
out[2,]=c(tau-out[1,1], rmst.se, tau-out[1,4], tau-out[1,3])
rownames(out)=c("RMST","RMTL")
colnames(out)=c("Est.", "se", paste("lower .",round((1-alpha)*100, digits=0), sep=""),
               paste("upper .",round((1-alpha)*100, digits=0), sep=""))

Z=list()
Z$result=out
Z$rmst = out[1,]
Z$rmtl = out[2,]
Z$tau=tau
Z$rmst.var = rmst.var
Z$fit=ft
class(Z)="rmst1_update"

return(Z)
}
```

We took the calculated `effDF` in the output of `rmst2_update.WS` and use it to pick a t-distribution quantile. Square of this quantile is then used to calculate the p-value for the log empirical likelihood.