Paper AS12

Assessing Non-Inferiority and Equivalence from Survival Data

Parveen Kumar, GenInvo, Mohali, India
Chintan Arora, GenInvo, Mohali, India

ABSTRACT

Non-inferiority or equivalence trials aim to show that the test drug is either not clinically worse than the reference (non-inferiority) or is clinically similar to the reference drug (equivalence) with margin calculation involved with both types of trials. There is ample literature available to asses non-inferiority or equivalence for continuous or categorical types of data. There are also copious options available in pre-defined procedures in statistical software like SAS® and R; however, there is little written about options available in these tools to assess non-inferiority and equivalence specifically for survival data or calculating sample size for this type of study design. SAS® has provided options to evaluate non-inferiority through PROC PHREG by calculating hazard ratio and corresponding CIs, but there are no options to evaluate non-inferiority if LOG-RANK TEST needs to be performed.

This paper will explore the procedures or libraries available in SAS® and R to calculate sample size and assess non-inferiority for survival data under various scenarios.

INTRODUCTION

Non-inferiority trials with time-to-event data as primary outcomes are becoming very common. Design and analysis of non-inferiority trials is challenging in many aspects such as selection of survival endpoint, calculation of non-inferiority margin, sample size estimation and data analysis. It becomes more challenging when there is little known about the endpoint, no historical data available, or standard tools do not have direct estimation methods.

The most common method for analysis of survival data is using survival distribution. In this context, the design and analysis of non-inferiority trials with time-to-event outcomes frequently relies on the hazard ratio (HR) as a measure of treatment effect. However, there are other estimates used by practitioners from time to time such as median survival, Kaplan Meir (KM) estimate, Restricted Mean Survival Time (RMST) and Restricted Mean Loss Time (RMLT). The primary survival endpoint should be decided at the start of the trial as this will determine the next steps in design and analysis of the non-inferiority trial.

ICH E10 has provided a guidance for selection of the control group and calculation of margin. However, there is dependency on published historical data for calculation of the non-inferiority margin. Hazard Ratio is a commonly used endpoint for survival distribution comparison between two groups, but it is based on the assumption of proportional hazard over the study period. If this assumption does not hold true based on historical data, then an assumption free measure is required. Choosing a different endpoint would also need a careful assessment of non-inferiority margin.

Sample Size estimation for survival endpoints is common and there are options available in standard software for this, but sample size estimation for non-inferiority assessment is more challenging as there are no direct options available in the same software. For example, SAS® PROC POWER provides options to do sample size estimation for non-inferiority comparisons for categorical and continuous endpoints but there are no direct options available for sample size estimation on survival data endpoints.

There are plenty of options available in SAS® or R software to analyze survival data with calculation of confidence intervals for treatment comparisons. However, there are no options to calculate p-value for non-inferiority testing. This paper will discuss challenges in design and analysis of survival data. The procedures available in SAS® and R to do sample size estimation and perform analysis following hazards model, median estimates or RMST for non-inferiority designed trials will also be discussed.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the authors. The interpretations of standards and procedures contained in this paper are those of the authors. Any views and recommendations stated within this paper are those of the authors, and they do not represent the position of their employers.

GENERAL HYPOTHESIS

In an equivalence trial, the statistical test aims at showing that two treatments are not too different in characteristics, where "not too different" is defined in a clinical manner. Equivalence is tested on both sides of estimates with specific margins. For example, in generic drugs, equivalence is tested by testing ratio of treatment groups is between 0.8 to 1.25. In a non-inferiority trial, the aim is to show that an experimental treatment is not (much) worse than a standard treatment where standard treatment may have lesser cost or comparatively better safety profile. In this case, margin will set on one side of interest.

NON-INFERIORITY HYPOTHESIS AND TESTING

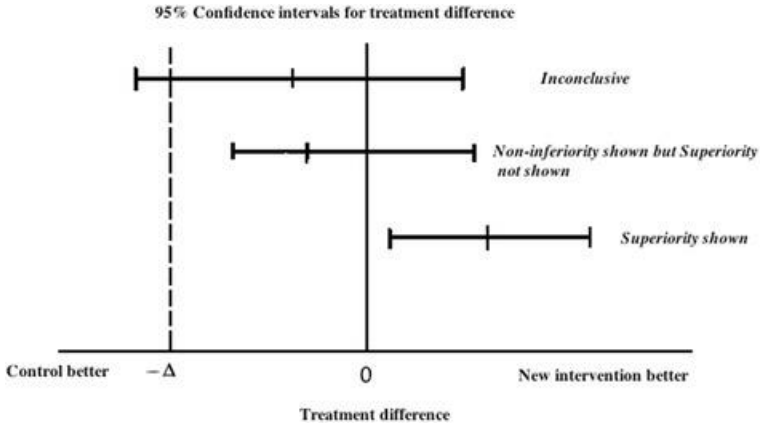
A general hypothesis is defined which can be adopted for the endpoint of interest. For comparison of estimate where higher value is better, hypothesis is defined as:

H0: Treatment comparison $\geq \delta_1$
HA: Treatment comparison $< \delta_1$

For comparison of estimate where lower value is better, hypothesis is defined as:

H0: Treatment comparison $\leq \delta 2$
HA: Treatment comparison $> \delta 2$

Non-inferiority can be concluded for above hypothesis using the confidence interval (CI) for the treatment comparison parameter analyzed as shown in the graph below:



Procedures available in SAS® such as PROC PHREG, PROC LIFETEST and packages in R such as ”survival” with survfit functions, etc., for survival analysis are being used to calculate confidence intervals for treatment comparison parameters. While these procedures provide p-value for global standard hypothesis testing, no specific options are available to support the same for non-inferiority testing with specified non-inferiority margin and corresponding p-value. Consequently, this has to be calculated using estimates and standard error obtained during analysis.

EQUIVALENCE HYPOTHESIS AND TESTING

Using the annotation used for non-inferiority testing above, the null and alternative hypotheses for equivalence testing are:

H0: $\delta_1 \leq$ Comparison estimate $\leq \delta_2$
HA: Comparison estimate $< \delta_1$ and Comparison estimate $> \delta_2$

The null hypothesis is made up of two simple one-sided hypotheses:

H01: Comparison estimate $\leq \delta_2$ versus HA1: Comparison estimate $> \delta_2$
H02: Comparison estimate $\geq \delta_1$ versus HA1: Comparison estimate $< \delta_1$

Schuirmann’s (1987) two one-sided tests (TOST) approach can be used to test equivalence. This equivalence test essentially reverses the roles of the usual null and alternative hypothesis. The null hypotheses can be tested at significance level α by constructing a $100(1 – 2\alpha)\%$ confidence interval, and not the usual $100(1 – \alpha)\%$. With upper bound of the $100(1 – 2\alpha)\%$ confidence interval less than δ_2 and the lower bound greater than δ_1, treatment and the control to be concluded as equivalent. This is the same as conducting two one-sided tests of the null hypotheses as done in non-inferiority testing. If both one-sided tests are rejected, one can conclude that the two groups are equivalent (their difference is confined within a small margin). P-value of the equivalence test is equal to the maximum of the probability levels of the two one-sided tests.

HR (Hazard Ratio) considering ratio between hazard rate of treatment and standard at certain time can be considered as interest of parameters. Considering L as lower bound and U as upper bound, below are null and alternative hypotheses:

H0: HR < L or HR > U (as two one-sided hypothesis H01: HR < L, H02: HR > U)
H1: L < HR < U

Equivalence testing can be performed in a similar way to non-inferiority testing by extending the approach to two one sided testing. Non-inferiority trials are also called one-sided equivalence trials. Hence, only methods to analyze survival data, calculation of margin and sample size estimation for non-inferiority are discussed.

CALCULATION OF MARGIN

As non-inferiority trials with time-to-event data are becoming increasingly common and may require larger sample size, choice of clinically relevant margin has to be dealt with. ICH E10 states that "An acceptable non-inferiority margin should be defined, taking into account the historical data and relevant clinical and statistical considerations". The margin chosen for a non-inferiority trial cannot be greater than the smallest effect size that the Control drug would be reliably expected to have compared with placebo in the setting of the planned trial. Smallest effect size of control with placebo comes from historical data of a similar trial. Non-inferiority margin shall be chosen to prove that active is not inferior to control but superior to placebo. The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment. It should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.

The 'historical' confidence interval compares the control treatment with placebo (c minus p). The planned trial comparing the active and control treatments will also produce a confidence interval (for a minus c). If these intervals are combined, an indirect confidence interval comparing the test product and placebo can be obtained (a minus p). Delta can be defined as the lower bound of a minus c that ensures that the lower bound of the indirect confidence interval of a minus p will be above zero. As the comparison is indirect it might be wise to be conservative and select some value smaller than that suggested by this indirect calculation. For example, the margin could be 50% or 25% of the control treatment effect compared to placebo.

For survival data, historical data would be needed for the endpoint of interest. For example, if the primary endpoint of interest is hazard rate, then hazard ratio of active treatment in comparison with placebo would be needed for margin calculation. Weir et al. (2018) have shown that Hazard Ratio values from historical data can be converted into RMST estimates for margin calculation and sample size estimation.

SAMPLE SIZE CALCULATION

The next step in conducting a trial would be to estimate a sample size for the non-inferiority trial with hypothesis defined, endpoint of interest selected, and margin calculated.

Median Survival/Hazard Ratio

Sample size calculated using median survival time can also be used for hazard ratio analysis using Cox Proportional Hazard models. Similarly, sample size calculated using Hazard Ratio estimates can be used for log rank analysis of two survival curves. Log Rank test and Cox Proportional Hazard model follows same fundamental. PROC POWER in SAS® does not provide any direct options to calculate sample size for non-inferiority hypothesis. However, it is easy to program in SAS® or R using the method suggested by Chow S et al. (2008) in their book, "Sample Size Calculations in Clinical Research":

$$n = \frac{1}{P1P2D}\left(\frac{Z_{1-\alpha}+Z_{1-\beta}}{\ln(HR1)-\ln(\delta)}\right)^2$$

Where P1 and P2 is allocation of proportion to two groups:

- n: total sample size
- D: overall probability of event occurring within study period, i.e., proportion of events from historical data
- ln(HR1): natural log of Hazard ratio from historical data
- ln(δ): natural log of non-inferiority margin hazard ratio
- α: level of significance
- β: type II error

There is also a package available in R, i.e., 'Trial Size', which can do the calculation for sample size same as above in single statement as below:

```
Cox.NIS(alpha, beta, loghr, p1, p2, d, margin)
```

PASS, sample size estimation software, does have an option to estimate sample size for non-inferiority and equivalence testing. It also uses the method suggest by Chow S et al. (2008).

Restricted Mean Survival Time (RMST)

Difference or ratio of restricted median survival time (RMST) has also emerged for time-to-event comparison as it does not get effected by violation of the proportional hazard assumption. Recently, Isabelle R Weir et al. (2018) have suggested that using the RMST in non-inferiority trials may result in decreased sample sizes while maintaining an adequate level of power. Sample size comparison based on Hazard Ratio and RMST is another topic for discussion not in the scope of this paper. Here, sample code considering R package 'SSRMST' is provided which has an option to calculate sample size for non-inferiority testing for RMST comparison.

```
a = ssrmst(ac_rate=ac_rate, ac_period=ac_period, tot_time=tot_time,
           tau=tau, scale0=scale0, scale1=scale1, margin=margin, ntest=20)
```

ac_rate:	Accrual rate: the number of patients per unit time
ac_period:	Accrual period: the time point at last accrual
ac_number:	Accrual number: the total number of accrual patients
tot_time:	Total study time: the time point at last follow-up
tau:	Truncation time point to calculate RMSTs
scale0, scale1:	Scale parameters for the Weibull distribution in both the control and the active treatment
margin:	Non-inferiority margin: a clinically acceptable difference in RMST

ANALYSIS OF SURVIVAL DATA

Survival Data is time-to-event data with right censoring. A clinical trial may be designed to test the non-inferiority of an active treatment over a control in regard to their survival distributions. As survival times are not normally distributed and are censored, proportional hazard regression is generally used to analyze data. LOGRANK test is generally used to test any difference between two survival curves and Cox proportional hazard model is used to compare hazard functions using treatment as factor. However, the formulation for testing the significance of Cox regression coefficient is identical to standard log-rank test and sample size formulas for one analysis can work for other. Proportional hazards assumption plays an important role in analysis using cox proportional hazard model. However, when two survival curves cross each other during assessment period, proportional hazards assumption can also be questionable or simply untenable.

There are handful of assumption free measures of the survival outcome, too. For example, the mean and median of event time can be used as summary statistics and the difference or ratio of these quantities can be used to estimate treatment effects. Median is favored due to censoring in survival data. One should also emphasize the importance of confidence interval on median survival times when survival analysis is being discussed. R Simon (1986) wrote about calculating a confidence interval on the ratio of median survivals when the survival distributions are exponential. Kallappa M. Koti (2012) suggested an effective way to design and analyze a non-inferiority or equivalence trial using ratio of median estimates. However, there could be challenges with these estimates if there are not enough events during trial period and mean or median is not estimable. The survival rate at a specified time is another commonly used summary statistic. Although it can be less problematic than the mean or median, it does not provide an overall summary of the time-to-event outcome.

The restricted mean survival time (RMST) is an alternative measure that is introduced by researchers which provides summary of whole survival curve up to time horizon, in contrast to survival rate at a specified time (Royston and Parmar 2013; Uno et al. 2014; Trinquart et al. 2016). The RMST has attracted practitioners for its straightforward interpretation and its capability to deal with nonproportional hazards. When two survival curves cross, the difference in the RMST between two groups still provides information about efficacy in a clinical trial, whereas the log-rank test fails to detect the significance and the hazard ratio becomes meaningless.

There are no options available in SAS® or R to calculate difference/ratio of median survival with confidence intervals. Hence, we will be discussing the methods of assessing non-inferiority for Hazard Ratio and RMST. Confidence intervals calculated based on hazard ratios can be used to assess equivalence and non-inferiority from survival data.

HAZARD RATIO

Assuming, overall survival is event of interest while comparing Test drug T with Control C, null hypothesis for non-inferiority testing of Hazard ratio can be defined as:

H0: HZ(T)/HZ(C) ≤ L vs
HA: HZ(T)/HZ(C) > L

Where HZ(T): Hazard rate for Test Drug T
HZ(C): Hazard rate for Control C

And L can be any value >1 taking into consideration acceptable margin for non-inferiority

PROC PHREG can be used to calculate Hazard Ratio of T versus C using following code:

```
proc phreg data=one;
class trt;
   model time*censor(0)=trt;
   hazardratio trt/cl=both;
run;
```

This will provide Hazard Ratio (HR) with p-value using Wald Chi-square statistic for global hypothesis, i.e., BETA=0, which is equivalent to testing Hazard Ratio = 1. Non-inferiority can be concluded using Wald's confidence interval using estimated confidence limits of HR. If upper confidence limit of HR is less than the non-inferiority margin then non-inferiority is established. However, p-value for non-inferiority testing is not coming directly from PROC PHREG and therefore must be calculated using outcome (Figure 1) of PHREG as below:

p-value for non-inferiority= PROBNORM[(estimate-LN(L))/SE)]
where estimate=Parameter estimate; SE= Standard Error

Figure 1:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	A	1	-0.67813	0.24591	7.6049	0.0058	0.508	trt A

Same can be obtained in R using survival package, coxph() as below:

```
library("survival")
coxfit <- coxph(Surv(time, status) ~ sex==1, data=lung)
coxfit
summary(coxfit)
```

ANALYSIS OF RMST

The restricted mean survival time (RMST) is an alternative treatment outcome measure that can be estimated as the area under the survival curve up to a prespecified time horizon and hence can account for all survival information before that time horizon. RMST is a well-established, yet underutilized measure that can be interpreted as the average event-free survival time up to a pre-specified, clinically important time point. The RMST difference means gain or loss in the event-free survival time due to active versus control during this period.

Let T be a nonnegative random variable that represents the failure time of an individual from a homogeneous population. The survivor function (also known as the survival function) of T is defined as:

$$S(t) = Pr(T>t)$$

Assume that tau is a prespecified time point of interest. Let R be the minimum of T and tau:

$$R = \min (T, tau)$$

The restricted mean survival time is defined as the expected value of R:

$$RMST(tau) = E(R) = E[\min(T,tau)]$$

It can be evaluated by the area under the survivor function over [0,tau] as:

$$RMST(tau) = \int S(u)du \text{ from } 0 \text{ to } tau$$

Non-Parametric analysis of RMST is available in SAS®/STAT 15.1 under PROC LIFTEST. RMST analysis focuses on non-parametric estimation and group comparison of RMST.

The null and alternative hypothesis to be tested here is:

H0: RMST1 – RMST2 ≥ -L
HA: RMST1 – RMST2 < -L

The following statements use PROC LIFETEST to perform analyses of the restricted mean survival time (RMST) in addition to the standard analyses:

```
proc lifetest data=one rmst(tau=40);
time time*event(0;
strata trt/diff=all ;
run;
```

Figure 2 shows the RMST analysis results from SAS®. It does not have Confidence Interval and p-value for non-inferiority which can be calculated using Z statistic and tabulated Z value as:

Confidence intervals = estimate ± Z(1-α/2)*SE
p-value for non-inferiority= PROBNORM[(estimate-LN(L))/SE)]
where estimate=Parameter estimate; SE= Standard Error

Figure 2:

Restricted Mean Survival Time Comparisons

Stratum Comparison		Difference	Standard Error	Chi-Square	Pr > ChiSq
A	B	10.79363	3.7997	8.0691	0.0045

There is also an R package available to compare two groups with respect to restricted mean survival time as:

```
library(survRM2)
rmst <- rmst2(time, status, arm, tau=, covariates=)
```

CONCLUSION

Design and analysis of non-inferiority testing for survival data has been discussed in this paper for three endpoints, i.e., Median survival, Hazard Ratio, and RMST. Analysis of median survival and hazard ratio follows a similar formulation and assumption, hence sample size estimation using median survival can be used for hazard ratio and vice versa. Non-inferiority margin calculation and sample size estimation is dependent on historical data and even if historical data is available in median survival and study is planned for hazard ratio, margin and sample size can be calculated for hazard ratio.

Treatment comparison for difference/ratio of median survival is not evolved enough to have direct options in SAS® or R to estimate confidence interval for treatment comparison. However, there has been research done to suggest new tests on non-inferiority assessment for difference or ratio of median survival time. However, non-inferiority assessment for Hazard Ratio and RMST in SAS® and R is straight forward. While not the focus of this paper, it should be noted that we had determined RMST to be a better measure than Hazard ratio whenever there is chance that proportional hazard assumption is not expected to be met.

REFERENCES

Kallappa M. Koti (2012). “New Tests for Assessing Non-Inferiority and Equivalence from Survival Data”. Open Journal of Statistics, 2013, 3, 55-64.

Royston, P., and Parmar, M. K. B. (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 13:152–166.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., et al. (2014). “Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis.” Journal of Clinical Oncology 32:2380–2385.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016). “Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials.” Journal of Clinical Oncology 34:1813–1819.

Isabelle R Weir and Ludovic Trinquart (2018). “Design of non-inferiority randomized trials using the difference in Restricted Mean Survival Times”. Clin Trials. 2018 Oct; 15(5): 499–508.

Chow S, Shao J, Wang H. 2008. Sample Size Calculations in Clinical Research. 2nd Ed. Chapman & Hall/CRC Biostatistics Series. page 177.

ACKNOWLEDGMENTS

I would like to thank Thomas Kalfas, Innovations Officer, GenInvo and Kalpesh Prajapati for thorough review of the paper and improving the quality of paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Parveen Kumar
GenInvo Inc.
702, Bestech Tower A, Mohali
Punjab, India - 160062
Work Phone: +91 9871819209
Email: parveen.k12@geninvo.com