Chapter 344

Multi-Arm Non-Inferiority Tests for Treatment and Control Survival Curves using Cox's Proportional Hazards Model

Introduction

This module computes power and sample size for multiple non-inferiority comparisons of treatment survival curves versus a control survival curve based on the results in Machin, Campbell, Tan, and Tan (2018). In this design, there are *k* treatment groups and one control group. A survival curve is measured in each group. A total of *k* non-inferiority hypothesis tests are anticipated each comparing a treatment group with the common control group using a simple z-test based on a Cox proportional hazards regression coefficient.

The formulation for testing the significance of a Cox regression coefficient is identical to the standard logrank test. Thus, the power and sample size formulas for one analysis also work for the other. The Cox Regression model has the added benefit over the exponential model that it does not assume that the hazard rates are constant, but only that they are proportional. That is, that the hazard ratio remains constant throughout the experiment, even if the hazard rates vary.

A Bonferroni adjustment of the type I error rate may be optionally made because several comparisons are being tested using the same data. Making a multiplicity adjustment is usually recommended, but not always. In fact, Saville (1990) advocates not applying it and Machin, Campbell, Tan, and Tan (2018) include omitting it as a possibility.

Background

Whether you want to test several doses of a single treatment or several types of treatments, good research practice requires that each treatment be compared with a control. For example, a popular three-arm design consists of three groups: control, treatment A, and treatment B. Two tests are run: treatment A versus control and treatment B versus the same control. This avoids having to obtain a second control group for treatment B. Besides the obvious efficiency in subjects, it may be easier to recruit subjects if their chances of receiving the new treatment are better than 50-50.

Technical Details

Cox's Proportional Hazards Regression

Cox's proportional hazards regression is widely used for survival data. The regression model is

$$h(t|z) = h(t|0)\exp(bz)$$

where

b is the regression coefficient which is equal to $\log[h(t|1)/h(t|0)] = \log(HR)$

z is a binary indicator variable of treatment group

t is elapsed time

$h(t|z)$ is the hazard rate at time t, given covariate z

HR is the hazard ratio, $h(t|1)/h(t|0)$

Non-Inferiority Hypothesis

Lower Hazards Better

Assuming that lower hazard rates are better, non-inferiority means that the treatment hazard rate is at most, only slightly higher than the control hazard rate. We find it more convenient to state the hypotheses in terms of the hazard ratio, *HR*, rather than the Cox regression coefficient, *b*. Remembering that *b* = log(*HR*) and assuming that $HR_0 > 1$, non-inferiority requires that $HR < HR_0$. Here, HR_0 is the boundary of clinical insignificance or the non-inferiority boundary.

The statistical hypotheses that results in the conclusion of non-inferiority when the null hypothesis is rejected is

$$H_0: \log(HR) \geq \log(HR_0) \quad \text{vs.} \quad H_a: \log(HR) < \log(HR_0)$$

Higher Hazards Better

Assuming that higher hazard rates are better, non-inferiority means that the treatment hazard rate is at most, only slightly lower than the control hazard rate. We find it more convenient to state the hypotheses in terms of the hazard ratio, *HR*, rather than the Cox regression coefficient, *b*. Remembering that *b* = log(*HR*) and assuming that $HR_0 < 1$, non-inferiority requires that $HR > HR_0$. Here, HR_0 is called the boundary of clinical insignificance or the non-inferiority boundary.

The statistical hypotheses that results in the conclusion of non-inferiority when the null hypothesis is rejected is

$$H_0: \log(HR) \leq \log(HR_0) \quad \text{vs.} \quad H_a: \log(HR) > \log(HR_0)$$

Test Statistic

It can be shown that the test of *b* based on the partial likelihood method of Cox (1972) coincides with the common logrank test statistic shown next.

Logrank Test

The logrank test statistic is

$$L = \frac{\sum_{k=1}^{K}\left(I_k - \frac{Y_{1i}HR_0}{Y_{1i}HR_0 + Y_{2i}}\right)}{\left[\sum_{k=1}^{K}\left(\frac{Y_{1i}Y_{2i}HR_0}{(Y_{1i}HR_0 + Y_{2i})^2}\right)\right]^{-\frac{1}{2}}}$$

where K is the number of deaths, Y_{ij} is the number of subjects at risk just prior to the j^{th} observed event in the i^{th} group, and I_k is a binary variable indicating whether the k^{th} event is from group 1 or not.

The distribution of L is approximately normal with mean $\left(\log(HR) - \log(HR_0)\right)\sqrt{P_1P_2dN}$ and unit variance, where

P_1 is the proportion of N that is in the control group

P_2 is the proportion of N that is in the treatment group

N is the total sample size

N_1 is the sample size from the control group, $N_1 = N(P_1)$

N_2 is the sample size from the treatment group, $N_2 = N(P_2)$

Pev_1 is probability of the event of interest in the control group

Pev_2 is probability of the event of interest in the treatment group

d is the overall probability of an event, $d = Pev_1P_1 + Pev_2P_2$

HR is the observed hazard ratio

HR_0 is the non-inferiority boundary (limit) of the hazard ratio

Power Calculations

The power of this test is given by

$$\Phi\left(\left(\log(HR_1) - \log(HR_0)\right)\sqrt{P_1P_2dN} - z_{1-\alpha}\right)$$

where HR_1 is the actual value of the hazard ratio under the alternative hypothesis.

Testing Multiple Treatment Groups versus a Single Control Group

Suppose you have *k* treatment groups with samples of size N_i and one control group with a sample of size N_C. The total sample size is $N = N_1 + N_2 + ... + N_k + N_C$. The response for each subject is their survival time until they either exhibit the event of interest or they are censored from the study.

A Cox proportional hazards regression model is fit to the data in which one of the independent variables is a binary variable that is zero if the subject is from the control group or one if they are from the i^{th} treatment group. Suppose that the regression coefficient associated with this independent variable is called b_i. As pointed out above, it turns out that

$$b_i = \log(HR_i)$$

where HR_i is the hazard ratio comparing the treatment and control groups. If HR_i = 1, there is no difference between the groups.

The data may be analyzed using *k* separate regressions each producing a non-inferiority test of the hazard ratio comparing a treatment group to the common control group.

The power for each of the *k* tests can be computed using the formula given above.

Multiplicity Adjustment

Because *k* z-tests between treatment groups and the control group are run when analyzing the results of this study, many statisticians recommend that the Bonferroni adjustment be applied. This adjustment is easy to apply: the value of alpha that is used in the test is found by dividing the original alpha by the number of tests. For example, if the original alpha is set at 0.05 and the number of treatment (not including the control) groups is five, the individual tests will be conducted using an alpha of 0.01.

The main criticism of this procedure is that if there are many tests, the value of alpha becomes very small. To mitigate against this complaint, some statisticians recommend separating the treatment groups into those that are of primary interest and those that are of secondary interest. The Bonferroni adjustment is made by the using the number of primary treatments rather than the total number of treatments.

There are some who advocate ignoring the adjustment entirely in the case of randomized clinical trials. See for example Saville (1990) and the discussion in chapter 14 of Machin, Campbell, Tan, and Tan (2018).

Size of the Control Group

Because the control group is used over and over, some advocate increasing the number of subjects in this group. The standard adjustment is to include $\sqrt{k}$ subjects in the control group for each subject in one of the treatment groups. See Machin, Campbell, Tan, and Tan (2018, pages 231-232). Note that often, the treatment groups all have the same size.

Example 1 – Finding the Sample Size

A parallel-group, non-inferiority trial is being designed to compare the survivability associated with three doses of a test compound against the standard (control) therapy in patients with a specific type of disease.

The proportion surviving one-year after the current treatment is 0.50 ($h_c = 0.693$). The researchers want to determine the sample size necessary to detect the situation when the proportion surviving one-year after the new treatment is 0.75 ($h_t = 0.288$). Hence, they want to compute the power when

$$HR = 0.288/0.693 = 0.4156$$

The non-inferiority hazard ratio is 1.25. The researchers would like to study the influence of *HR* on the sample size, so they would like to look at a range of possible values: 0.4 to 1.

For planning purposes, they decide that the probability of an event is 0.50 in the control group and 0.25 in the three treatment groups. The researchers decide to use a 0.025 significance level and a power of 0.8.

Following standard procedure, the control group multiplier will be set to $\sqrt{k} = \sqrt{3} = 1.732$ since the control group is used for three comparisons in this design.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Sample Size**
Higher Hazards Are	**Worse (H1: HR < HR0)**
Power of Each Test	**0.80**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Enter Group Allocation Pattern, solve for group sample sizes**
Pev (Default Probability of an Event)	**0.75**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**
Control Probability of an Event	**0.5**
Control Sample Size Allocation	**1.732**
Set A Number of Groups	**3**
Set A Hazard Ratio	**0.4 0.6 0.8 1**
Set A Probability of an Event	**0.25**
Set A Sample Size Allocation	**1**
Set B Number of Groups	**0**
Set C Number of Groups	**0**
Set D Number of Groups	**0**
More	**Unchecked**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Reports

Numeric Results

Solve For:	Sample Size
Group Allocation:	Enter Group Allocation Pattern, solve for group sample sizes
Test Type:	Z-Test Based on the Log Hazard Ratio
Hypotheses:	H0: HR ≥ HR0 vs. H1: HR < HR0
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power		Sample Size			Hazard Ratio			Alpha	
Comparison	Target	Actual	Ni	Allocation	Events Ei	Actual HRi	NI Limit HR0	Probability of an Event Pevi	Overall	Bonferroni-Adjusted
Control			55	1.732	27.5			0.50		
vs A1	0.8	0.81050	32	1.000	8.0	0.4	1.25	0.25	0.025	0.00833
vs A2	0.8	0.81050	32	1.000	8.0	0.4	1.25	0.25	0.025	0.00833
vs A3	0.8	0.81050	32	1.000	8.0	0.4	1.25	0.25	0.025	0.00833
Total			151		51.5					
Control			132	1.732	66.0			0.50		
vs A1	0.8	0.80635	76	1.000	19.0	0.6	1.25	0.25	0.025	0.00833
vs A2	0.8	0.80635	76	1.000	19.0	0.6	1.25	0.25	0.025	0.00833
vs A3	0.8	0.80635	76	1.000	19.0	0.6	1.25	0.25	0.025	0.00833
Total			360		123.0					
Control			352	1.732	176.0			0.50		
vs A1	0.8	0.80033	203	1.000	50.8	0.8	1.25	0.25	0.025	0.00833
vs A2	0.8	0.80033	203	1.000	50.8	0.8	1.25	0.25	0.025	0.00833
vs A3	0.8	0.80033	203	1.000	50.8	0.8	1.25	0.25	0.025	0.00833
Total			961		328.3					
Control			1406	1.732	703.0			0.50		
vs A1	0.8	0.80001	812	1.000	203.0	1.0	1.25	0.25	0.025	0.00833
vs A2	0.8	0.80001	812	1.000	203.0	1.0	1.25	0.25	0.025	0.00833
vs A3	0.8	0.80001	812	1.000	203.0	1.0	1.25	0.25	0.025	0.00833
Total			3842		1312.0					

Comparison	The group that is involved in the comparison between the treatment and control displayed on this report line. The comparison is made using the hazard ratio.
Target Power	The power desired. Power is probability of rejecting a false null hypothesis for this comparison. This power is of the comparison shown on this line only.
Actual Power	The power actually achieved.
Ni	Sample Size. The number of subjects in the ith group. The total sample size, N, is shown as the last row of the column.
Allocation	The group sample size allocation pattern. The value on each row represents the relative number of subjects assigned to the group.
Ei	The number of events in the ith group required to achieve the power indicated. Ei = Pevi × Ni.
HRi	The hazard ratio of the ith treatment group. HR = hi / hc.
HR0	The non-inferiority hazard ratio boundary used to declare whether a treatment is non-inferior to the control.
Pevi	The average probability that a subject the ith group will have an event during the study. Pevi also represents the proportion of individuals in the ith group that are expected to have an event during the study. This probability includes the impact of various kinds of censoring.
Overall Alpha	The probability of rejecting at least one of the comparisons in this experiment when each null hypothesis is true.
Bonferroni Alpha	The adjusted significance level at which each individual comparison is made.

Summary Statements

A parallel, 4-group design (with one control group and 3 treatment groups) will be used to test whether the hazard rate for each treatment group is non-inferior to the control group hazard rate, with a non-inferiority hazard ratio of 1.25 (H0: HR ≥ 1.25 versus H1: HR < 1.25, HR = treatment hazard rate i / control hazard rate). In this study, higher hazard rates are considered to be worse. The non-inferiority hypotheses will be evaluated using 3 one-sided, two-sample, Bonferroni-adjusted, Cox's proportional hazards regression term Z-tests, with an overall (experiment-wise) Type I error rate (α) of 0.025. It is anticipated that the proportions of subjects in each group that will have an event during the course of the study (beginning with the control group) will be 0.5, 0.25, 0.25, and 0.25. To detect the treatment to control hazard ratios 0.4, 0.4, and 0.4 with at least 80% power for each test, the control group sample size needed will be 55 and the number of needed subjects for the treatment groups will be 32, 32, and 32 (totaling 151 subjects overall). The corresponding total number of events is 51.5. These results assume that the hazard ratios are constant throughout the study.

Dropout-Inflated Sample Size

Group	Dropout Rate	Sample Size Ni	Dropout-Inflated Enrollment Sample Size Ni'	Expected Number of Dropouts Di
1	20%	55	69	14
2	20%	32	40	8
3	20%	32	40	8
4	20%	32	40	8
Total		151	189	38
1	20%	132	165	33
2	20%	76	95	19
3	20%	76	95	19
4	20%	76	95	19
Total		360	450	90
1	20%	352	440	88
2	20%	203	254	51
3	20%	203	254	51
4	20%	203	254	51
Total		961	1202	241
1	20%	1406	1758	352
2	20%	812	1015	203
3	20%	812	1015	203
4	20%	812	1015	203
Total		3842	4803	961

Group	Lists the group numbers.
Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
Ni	The evaluable sample size for each group at which power is computed (as entered by the user). If Ni subjects are evaluated out of the Ni' subjects that are enrolled in the study, the design will achieve the stated power.
Ni'	The number of subjects that should be enrolled in each group in order to obtain Ni evaluable subjects, based on the assumed dropout rate. Ni' is calculated by inflating Ni using the formula Ni' = Ni / (1 - DR), with Ni' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
Di	The expected number of dropouts in each group. Di = Ni' - Ni.

Dropout Summary Statements

Anticipating a 20% dropout rate, group sizes of 69, 40, 40, and 40 subjects should be enrolled to obtain final group sample sizes of 55, 32, 32, and 32 subjects.

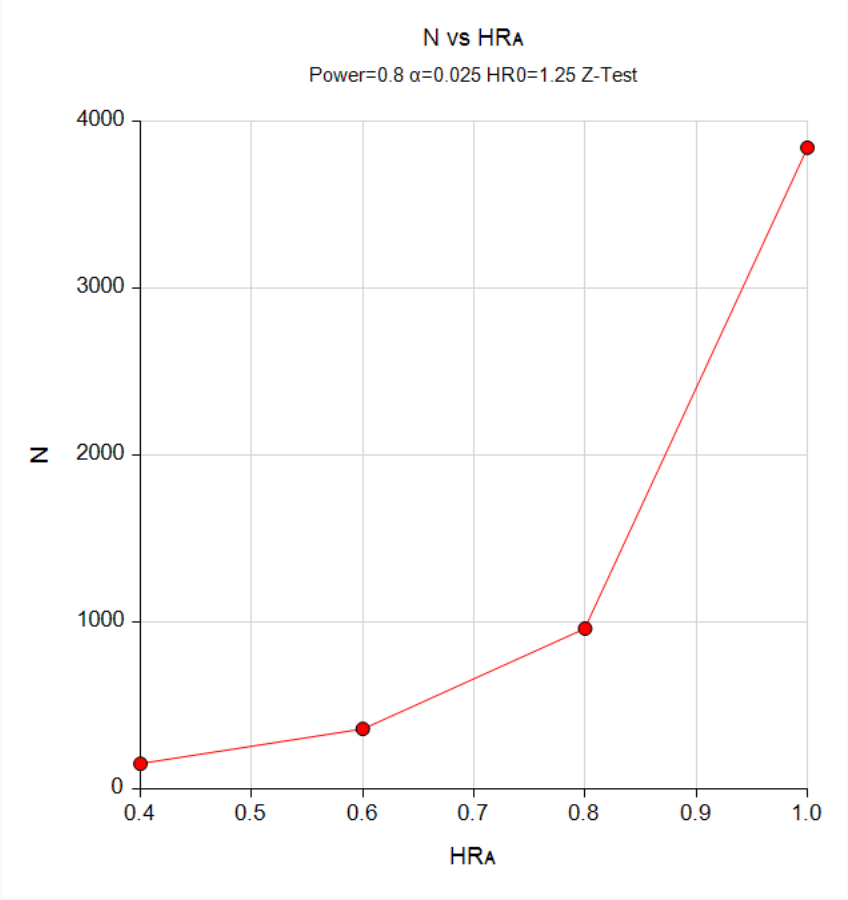
References

Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.
Machin, D., Campbell, M.J., Tan, S.B, and Tan, S.H. 2018. Sample Sizes for Clinical, Laboratory, and Epidemiology Studies, 4th Edition. Wiley Blackwell.
Schoenfeld, David A. 1983. 'Sample Size Formula for the Proportional-Hazards Regression Model', Biometrics, Volume 39, Pages 499-503.

This report shows the numeric results of this power study. Notice that the results are shown in blocks of three rows at a time. Each block represents a single design.

Plots Section

Plots



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the sample size of changing the hazard ratio from 0.4 to 1.0.

Example 2 – Validation using a Previously Validated Procedure

We could not find a validation result in the statistical literature, so we will use a previously validated **PASS** procedure (**Non-Inferiority Tests for Two Survival Curves Using Cox's Proportional Hazards Model**) to produce the results for the following example.

A parallel-group clinical trial is being designed to compare the survivability induced by three doses of a test compound against the standard (control) therapy in patients with a specific type of disease.

The hazard ratio for the power calculation is set to 1.0.

For planning purposes, they decide that the probability of an event is 0.50 in the control group and 0.25 in the three treatment groups. The researchers decide on a 0.025 significance level and a power of 0.8. Since a Bonferroni adjustment is made, the significance level is reduced to 0.025 / 3 = 0.00833.

The sample sizes of all groups will be equal.

The **Non-Inferiority Tests for Two Survival Curves Using Cox's Proportional Hazards Model** procedure is set up as follows.

Design Tab

Solve For	**Sample Size**
Higher Hazards Are	**Worse (Ha: HR < HR0)**
Power	**0.80**
Alpha	**0.008333** (which is Alpha / k)
Group Allocation	**Equal (N1 = N2)**
Pev1 (Event Probability in Group 1)	**0.5**
Pev2 (Event Probability in Group 2)	**0.25**
HR1 (Actual Hazard Ratio)	**1**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**

This set of options generates the following report.

Numeric Results

Solve For: Sample Size
Hypotheses: H0: HR ≥ HR0 vs. Ha: HR < HR0

Power	Total Sample Size N	Control Sample Size N1	Trtmnt Sample Size N2	Percent Control %N1	Actual Hazard Ratio HR1	Non-Inf Hazard Ratio HR0	Control Prob Event Pev1	Trtmnt Prob Event Pev2	Control Events E1	Trtmnt Events E2	Alpha
0.8003	2244	1122	1122	50	1	1.25	0.5	0.25	561	280.5	0.008

In order to maintain a power of 80% for all three groups, it is apparent that the groups will all need to have a sample size of 1122 per group. This table contains the validation values. We will now run these values through the current procedure and compare the results with these values.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Sample Size**
Higher Hazards Are	**Worse (H1: HR < HR0)**
Power of Each Test	**0.80**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Nc = N1 = N2 = ...)**
Pev (Default Probability of an Event)	**0.75**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**
Control Probability of an Event	**0.5**
Set A Number of Groups	**3**
Set A Hazard Ratio	**1**
Set A Probability of an Event	**0.25**
Set B Number of Groups	**0**
Set C Number of Groups	**0**
Set D Number of Groups	**0**
More	**Unchecked**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Group Allocation:	Equal (Nc = N1 = N2 = ...)
Test Type:	Z-Test Based on the Log Hazard Ratio
Hypotheses:	H0: HR ≥ HR0 vs. H1: HR < HR0
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power				Hazard Ratio			Alpha	
Comparison	Target	Actual	Sample Size Ni	Events Ei	Actual HRi	NI Limit HR0	Probability of an Event Pevi	Overall	Bonferroni-Adjusted
Control			1122	561.0			0.50		
vs A1	0.8	0.80026	1122	280.5	1	1.25	0.25	0.025	0.00833
vs A2	0.8	0.80026	1122	280.5	1	1.25	0.25	0.025	0.00833
vs A3	0.8	0.80026	1122	280.5	1	1.25	0.25	0.025	0.00833
Total			4488	1402.5					

As you can see, the sample sizes are all 1122. This matches the sample size found in the validation run above. The procedure is validated.