Chapter 627

Multi-Arm Non-Inferiority Tests for Survival Curves using Cox's Proportional Hazards Model in a Cluster-Randomized Design

Introduction

This procedure calculates power and sample size for testing the non-inferiority of multiple treatment hazard rates versus a common control hazard rate using Cox's proportional hazards regression when the data are obtained in a cluster-randomized design. Because survival times are not normally distributed and because some survival times are censored, Cox proportional-hazards regression is often used to analyze the data. 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 procedure is documented in Chow, Shao, Wang, and Lokhnygina (2018) and Machin, Campbell, Tan, and Tan (2018) which are based on the work of Schoenfeld (1981, 1983).

A *cluster (group) randomized design* is one in which whole units, or clusters, of subjects are randomized to the groups rather than the individual subjects in those clusters. However, the conclusions of the study concern individual subjects rather than the clusters. Examples of clusters are families, school classes, neighborhoods, hospitals, and doctor's practices.

Cluster-randomized designs are often adopted when there is a high risk of contamination if cluster members were randomized individually. For example, it may be difficult for doctors to use two treatment methods in their practice. The price of randomizing by clusters is a loss of efficiency--the number of subjects needed to obtain a certain level of precision in a cluster-randomized trial is usually much larger than the number needed when the subjects are randomized individually. Hence, standard methods of sample size estimation cannot be used.

The 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.

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 $h(t|1)/h(t|0)$

The one-sided, statistical hypothesis testing survival non-inferiority is a test of whether *b* is different from a constant value. This hypothesis is stated as

$$H_0: b \geq b_0 \quad \text{vs.} \quad H_1: b < b_0$$

Non-Inferiority Hypothesis

Higher Hazards Worse

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 result in the conclusion of non-inferiority when the null hypothesis is rejected are of the form

$$H_0: \log(HR) \geq \log(HR_0) \quad \text{vs.} \quad H_1: \log(HR) < \log(HR_0)$$

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 result in the conclusion of non-inferiority when the null hypothesis is rejected are of the form

$$H_0: \log(HR) \leq \log(HR_0) \quad \text{vs.} \quad H_1: \log(HR) > \log(HR_0)$$

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

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

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 $\left(\log(HR_i) - \log(HR_0)\right)\sqrt{P_C P_i d_i N}$ and unit variance, where

P_C is the proportion of *N* that is in the control group

P_i is the proportion of *N* that is in the i^{th} treatment group

N is the total sample size

N_C is the sample size from the control group, $N_C = N(P_C)$

N_i is the sample size from the i^{th} treatment group, $N_i = N(P_i)$

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

Pev_i is probability of the event of interest in the i^{th} treatment group

d_i is the overall probability of an event, $d_i = Pev_C P_C + Pev_i P_i$

HR_i is the observed hazard ratio for the i^{th} treatment group vs. the control group

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

Cluster-Randomized Designs

Denote an observation by Y_{ijk} where i = c,1, 2, ..., G gives the group, j = 1, 2, ..., K_i gives the cluster within group i, and k = 1, 2, ..., m_{ij} denotes an individual in cluster j of group i. In this chapter, we will assume that group c is the control group and groups 1, ..., G are the treatment groups.

Let ρ denote the intracluster correlation coefficient (ICC) among individuals from the same cluster. This correlation is the correlation of censor indicator variable. Let COV denote the coefficient of variation of the cluster sizes. Machin *et al*. (2018) page 101 shows that the number of events, e, that are needed to obtain a given power of 1 – β and a significance level of α to detect a hazard ratio of HR_i (h_i / h_C) is given by

$$e = \frac{(1+r)^2}{r}\left[\frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{(\log HR_i - \log HR_0)^2}\right]$$

where $r = N_i / N_C$ and $z_x = \Phi(x)$ is the standard normal distribution function.

The number of subjects in a regular design can be determined using

$$N = N_C + N_i = \frac{e}{F}$$

where

$$F = \frac{(Pev_C + r(Pev_i))}{1+r}$$

Design Effect

When using a cluster-randomized design Machin *et al*. (2018) page 202 show that the above formula is modified using a quantity that is known as the *design effect* (DE). The version of DE that is used in **PASS** is given as formula 12.7 on page 197 of Machin *et al*. (2018) which is

$$DE = 1 + \{[CV^2 + 1]\bar{M} - 1\}\rho$$

where $\bar{M}$ is the average cluster size of all clusters given by

$$\bar{M} = \frac{K_C M_C + K_i M_i}{K_C + K_i}$$

CV is the coefficient of variation of cluster sizes of all clusters in the study and ρ is the ICC as defined above.

The resulting sample size formula in terms of cluster counts and size is

$$N = K_C M_C + K_i M_i = DE\left(\frac{e}{F}\right)$$

Power Calculations

The power of the one-sided, statistical test of b is given by

$$\Phi\left(b\sqrt{P_C P_i d_i N} - z_{1-\alpha}\right)$$

or equivalently

$$\Phi\left(\log(HR_i)\sqrt{P_C P_i d_i N} - z_{1-\alpha}\right)$$

where HR_i is the actual assumed value of the hazard ratio for treatment group *i* under the alternative hypothesis and N is defined as above.

Multiplicity Adjustment

Because *G* tests between the treatment groups and the control group are run when analyzing the results of this study, many statisticians recommend that a 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 should 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 clusters in this group. The standard adjustment is to include $\sqrt{G}$ clusters in the control group for each cluster 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 sample size.

Example 1 – Finding the Sample Size

Suppose a four-arm, cluster-randomized study is to be conducted in which higher hazards are worse; *HR0* = 1.25; $HR_1 = HR_2 = HR_3 = 1$, ρ = 0.01, $Pev_1 = Pev_2 = Pev_3 = 0.61$, $Pev_C = 0.82$, *Mi* = 10, 20, or 30, *COV* = 0.65, *alpha* = 0.025, and number of clusters is to be calculated. The target power is 0.9 calculated for a one-sided test.

The control group cluster allocation will be set to $\sqrt{G} = \sqrt{3} = 1.732$ 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	**0.90**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Enter Group Allocation Pattern, solve for group numbers of clusters**
M (Average Cluster Size)	**10 20 30**
COV of Cluster Sizes	**0.65**
Pev (Default Probability of an Event)	**0.75**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**
Control Probability of an Event	**0.82**
Control Average Cluster Size	**M**
Control Cluster Allocation	**1.732**
Set A Number of Groups	**3**
Set A Hazard Ratio	**1**
Set A Probability of an Event	**0.61**
Set A Average Cluster Size	**M**
Set A Cluster Allocation	**1**
Set B Number of Groups	**0**
Set C Number of Groups	**0**
Set D Number of Groups	**0**
More	**Unchecked**
ρ (Intracluster Correlation)	**0.01**

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 numbers of clusters
Higher Hazards Are:	Worse
Hypotheses:	H0: HR ≥ HR0 vs. H1: HR < HR0
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

		Sample	Number of			Cluster Size			Hazard Ratio				Alpha	
Comparison	Power	Size Ni	Events Ei	Clusters Ki	Cluster Allocation	Average Mi	COV	Probability of an Event Pevi	Non-Inferiority HR0	Actual HRi	ICC ρ	Design Effect DE	Overall	Bonferroni-Adjusted
Control		1140	1058.4	114	1.732	10	0.65	0.82			0.01	1.13225		
vs A1	0.90349	660	455.8	66	1.000	10	0.65	1.25	0.61	1	0.01	1.13225	0.025	0.00833
vs A2	0.90349	660	455.8	66	1.000	10	0.65	1.25	0.61	1	0.01	1.13225	0.025	0.00833
vs A3	0.90349	660	455.8	66	1.000	10	0.65	1.25	0.61	1	0.01	1.13225	0.025	0.00833
Total		3120	2426.0	312										
Control		1280	1337.7	64	1.732	20	0.65	0.82			0.01	1.27450		
vs A1	0.90244	740	575.3	37	1.000	20	0.65	1.25	0.61	1	0.01	1.27450	0.025	0.00833
vs A2	0.90244	740	575.3	37	1.000	20	0.65	1.25	0.61	1	0.01	1.27450	0.025	0.00833
vs A3	0.90244	740	575.3	37	1.000	20	0.65	1.25	0.61	1	0.01	1.27450	0.025	0.00833
Total		3500	3063.6	175										
Control		1440	1672.9	48	1.732	30	0.65	0.82			0.01	1.41675		
vs A1	0.90777	840	725.9	28	1.000	30	0.65	1.25	0.61	1	0.01	1.41675	0.025	0.00833
vs A2	0.90777	840	725.9	28	1.000	30	0.65	1.25	0.61	1	0.01	1.41675	0.025	0.00833
vs A3	0.90777	840	725.9	28	1.000	30	0.65	1.25	0.61	1	0.01	1.41675	0.025	0.00833
Total		3960	3850.7	132										

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.
Power	The probability of rejecting a false null hypothesis for this comparison. This power is of the comparison shown on this line only.
Ni	The number of items in the ith group. The total sample size is shown as the last row of the column.
Ei	The number of events in the ith group required to achieve the power indicated. Ei = Pevi × Ni.
Ki	The number of clusters in the ith group. The total number of clusters is reported in the last row of the column.
Allocation	The cluster allocation ratio of the ith group. The value on each row represents the relative number of clusters assigned to the group.
Mi	The average number of items per cluster (or average cluster size) in the ith group.
COV	The coefficient of variation of the cluster sizes within the group.
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.
HR0	The non-inferiority hazard ratio boundary used to declare whether a treatment is non-inferior to the control.
HRi	The hazard ratio of the ith treatment group. HR = hi / hc.
ρ	The intracluster correlation (ICC). The correlation between subjects within a cluster.
DE	The design effect. This value is used to increase the sample size because of the cluster randomization that is used in the design.
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 cluster-randomized design (with one control group and 3 treatment groups, and where higher hazard rates are considered worse) 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: HRi ≥ 1.25 versus H1: HRi < 1.25, HRi = hi / hc). The hypotheses will be evaluated using 3 one-sided, two-sample, Bonferroni-adjusted (divisor = 3) Cox regression coefficient tests, with an overall (experiment-wise) Type I error rate (α) of 0.025. The coefficient of variation of the cluster sizes in all clusters is assumed to be 0.65. The average probability of an event for a subject in the control group is assumed to be 0.82, and the event probabilities for the treatment groups are assumed to be 0.61, 0.61, and 0.61. The calculations are based on the assumption that the hazard ratio is constant throughout the study. The intracluster correlation is assumed to be 0.01. The average cluster size (number of subjects or items per cluster) for the control group is assumed to be 10, and the average cluster size for each of the treatment groups is assumed to be 10, 10, and 10. To detect the treatment to control hazard ratios 1, 1, and 1 with at least 90% power for each test, the control group cluster count needed will be 114 and the number of needed clusters for the treatment groups will be 66, 66, and 66 (totaling 312 clusters overall).

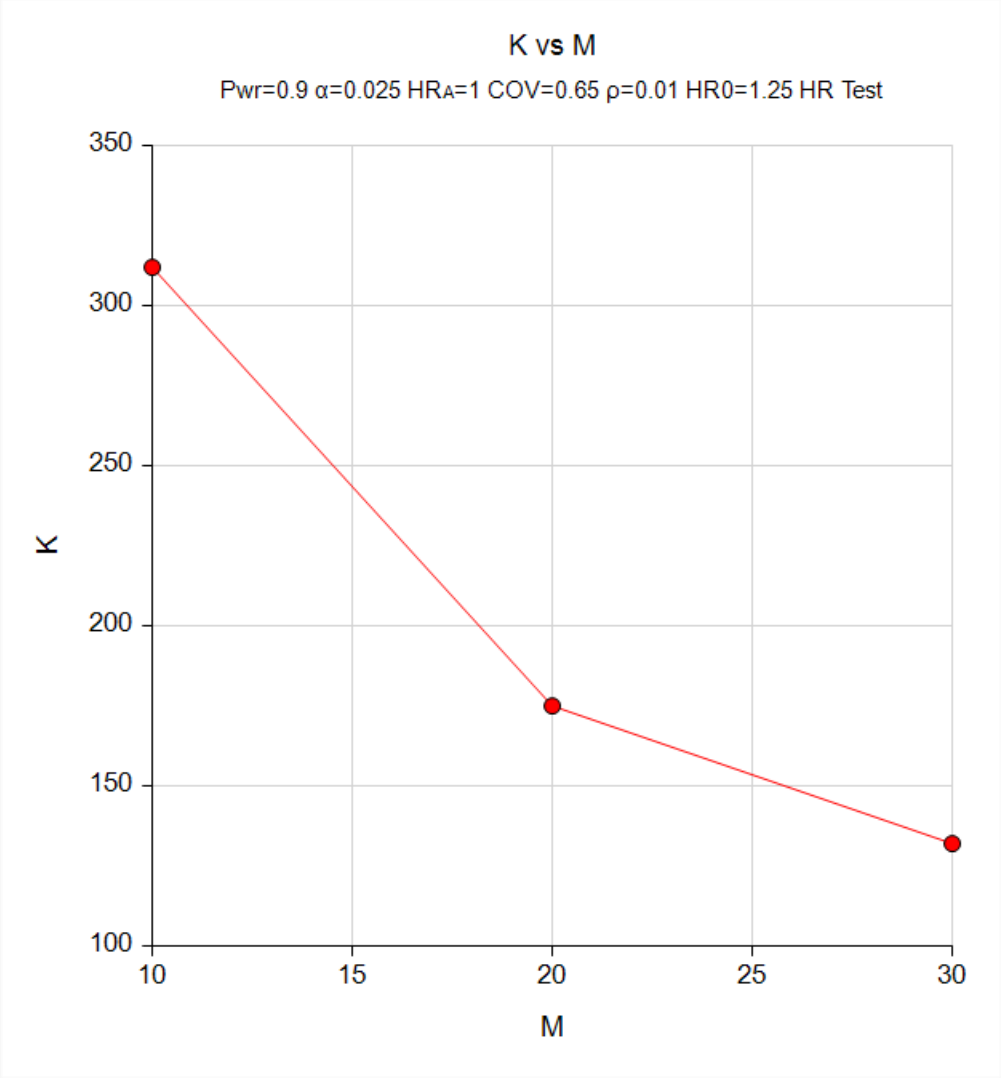
References

Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.
Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley. New York.
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, Third Edition. Taylor & Francis/CRC. Boca Raton, Florida.
Donner, A. and Klar, N. 1996. 'Statistical Considerations in the Design and Analysis of Community Intervention Trials'. J. Clin. Epidemiol. Vol 49, No. 4, pages 435-439.
Donner, A. and Klar, N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold. London.
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 sample size study. Notice that the results are shown in blocks of five rows at a time. Each block represents an individual treatment.

Plots Section

Plots



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the total cluster count, K, of increasing the average cluster size, M.

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 in a Cluster-Randomized Design**) to produce the results for the following example.

Suppose that a two-arm, cluster-randomized study is to be conducted in which $HR_1 = HR_2 = 1, HR0 = 1.25$, $\rho = 0.05$, $Pev_1 = Pev_2 = 0.7$, $Pev_C = 0.8$, $Mi = 2$, $COV = 0.6$, $alpha = 0.0125$, and number of clusters is 200 in each group. The resulting power is 0.63106 for a one-sided test.

The **Non-Inferiority Tests for Two Survival Curves using Cox's Proportional Hazards Model in a Cluster-Randomized Design** procedure is set up as follows.

Design Tab

Solve For	**Power**
Higher Hazards Are	**Worse (H1: HR < HR0)**
Alpha	**0.0125**
K1 (Number of Clusters)	**200**
M1 (Average Cluster Size)	**2**
K2 (Number of Clusters)	**K1**
M2 (Average Cluster Size)	**M1**
COV of Cluster Sizes	**0.6**
Pev1 (Probability of a Control Event)	**0.8**
Pev2 (Probability of a Treatment Event)	**0.7**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**
HR1 (Actual Hazard Ratio)	**1**
ρ (Intracluster Correlation)	**0.05**

This set of options generates the following report.

Numeric Results

Solve For: Power
Higher Hazards Are: Worse
Groups: 1 = Control, 2 = Treatment
Hypotheses: H0: HR ≥ HR0 vs. H1: HR < HR0

					Cluster Size			Hazard Ratio				
Group	Power	Sample Size Ni	Number of Events Ei	Number of Clusters Ki	Average Mi	COV	Probability of an Event Pevi	Non-Inferiority HR0	Actual HR1	Intracluster Correlation ρ	Alpha
1: Control		400	347.5	200	2	0.6	0.8			0.05	
2: Treatment	0.64843	400	304.1	200	2	0.6	0.7	1.25	1	0.05	0.0125
Total		800	651.6	400							

The power is computed to be 0.64843.

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	**Power**
Higher Hazards Are	**Worse (H1: HR < HR0)**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Kc = K1 = K2 = ...)**
Ki (Group Number of Clusters)	**200**
M (Average Cluster Size)	**2**
COV of Cluster Sizes	**0.6**
Pev (Default Probability of an Event)	**0.75**
HR0 (Non-Inferiority Hazard Ratio)	**1.25**
Control Probability of an Event	**0.8**
Control Average Cluster Size	**M**
Set A Number of Groups	**2**
Set A Hazard Ratio	**1**
Set A Probability of an Event	**0.7**
Set A Average Cluster Size	**M**
Set B Number of Groups	**0**
Set C Number of Groups	**0**
Set D Number of Groups	**0**
More	**Unchecked**
ρ (Intracluster Correlation)	**0.05**

Output

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

Numeric Results

Solve For:	Power
Higher Hazards Are:	Worse
Hypotheses:	H0: HR ≥ HR0 vs. H1: HR < HR0
Number of Groups:	3
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 2)

			Number of		Cluster Size			Hazard Ratio				Alpha	
Comparison	Power	Sample Size Ni	Events Ei	Clusters Ki	Average Mi	COV	Probability of an Event Pevi	Non-Inferiority HR0	Actual HRi	ICC ρ	Design Effect DE	Overall	Bonferroni-Adjusted
Control		400	347.5	200	2	0.6	0.8			0.05	1.086		
vs A1	0.64843	400	304.1	200	2	0.6	0.7	1.25	1	0.05	1.086	0.025	0.0125
vs A2	0.64843	400	304.1	200	2	0.6	0.7	1.25	1	0.05	1.086	0.025	0.0125
Total		1200	955.7	600									

As you can see, the power is 0.64843 for both treatment groups which matches the power found in the validation run above. The procedure is validated.