Downloadable SAS Programs

Sample Size and Power for a Logrank Test and Cox Proportional Hazards Model with Multiple Groups and Strata, or a Quantitative Covariate with Multiple Strata

John M. Lachin

The Biostatistics Center
Departments of Epidemiology and Biostatistics, and Statistics
The George Washington University
6110 Executive Boulevard, Suite 750
Rockville, Maryland USA 20852

This document describes the programs that are available for download that performed the calculations presented in the paper:

> Lachin JM. Sample size and power for a logrank test and Cox proportional hazards model with multiple groups and strata, or a quantitative covariate with multiple strata. *Statistics in Medicine* 2013; 32:4413–4425. DOI: 10.1002/sim.5839.

A brief description of each program follows. Note that these are not general use programs and each will have to be customized for each application.

Expwrlogl z: computes the power for a study with staggered entry and non-linear recruitment using the method of Lachin and Foulkes (1986). \
4gpexSSlogl z: Computes the sample size for the K-group test.
4gpexpwlogl z: Computes the power of the K-group test.
3,7 exsslogl z rev 4: computes sample size for pairwise comparisons.
4 gpstratPWlogrank: computes power for a stratified 4-group comparison.
PWQuantPH: computes the Power for a quantitative covariate in PH model
4gpPWQuantInt: computes Power for test of K-group by quantitative covariate interaction

To illustrate the use of each program, the following is the text from the example in the paper that shows which of the programs was used for the different computations. The text of the example includes some latex fragments but the program usage should otherwise be clear.

Note that there are two typos in the published paper. These are shown in the text below using track changes.

Example

3,7 exsslogl z.sas was used in the following text.

The Glycemia Reduction Approaches in Diabetes, A Comparative Effectiveness Study (GRADE) is designed to compare the effectiveness of four classes of drugs commonly used for treatment of type 2 diabetes. The primary outcome is the time to confirmed inability to maintain adequate glycemic control, which from prior studies is estimated to have a reference hazard rate of \lambda =0.0875 per year in the drug group(s) with the least durable effect on glycemic control. Herein we describe sample size and power computations assuming an R=3 year recruitment interval with a total duration of T=7 years. To allow for a lag in recruitment it is assumed that 40% of subjects are recruited in the first half of the recruitment period, 60% in the second, that corresponds to a

recruitment shape parameter of \gamma =-0.27. This yields a mean recruitment time of 1.7 years and a mean potential exposure of 4.8 years assuming no losses to follow-up. Allowing for 4% losses per year (\eta =0.04), from (\ref{Ef) the mean exposure time is reduced to 3.84 years with an event probability of 0.335 and a loss-to-follow-up probability of 0.153.

4gpexsslogl z was used in the following text. The second output for the first subsequent paragraph, the first output for the second.

A single overall global test of the hypothesis of equality among the 4 groups will be conducted on 3 df. The simplest alternative hypothesis is that three treatments all have the reference hazard rate of 0.0875 and one treatment (say the first) is superior to the other three with a hazard ratio of 0.75 versus the others, i.e. with a hazard of 0.0656 for the first and 0.0875 for the other three groups, and the vector of hazard ratios \beta =(0.75 1 1)^{T with group 4 as the reference. With equal sized treatment groups (\xi _{j}=1/4), then the expected probabilities of the event are 0.265 in the first and 0.335 in each of the other 3 groups. These yield a weighted mean log hazard \theta =-2.496 corresponding to a geometric mean hazard of 0.0824 and a non-centrality factor of \phi ^{2}=0.004338. The non-centrality parameter value that provides 90% power for a 3 df test at the 0.05 level is \psi ^{2(0.05,0.10,3)=14.1715. Substituting into (\ref{N), N=\psi ^{2} / \phi ^{2}=3268 (rounded up from 3267) would be required to provide 90% power to detect the hazard ratio of 0.75 for one therapy versus the others. This yields 216 subjects expected to have the event in the first group and 274 in each of the other 3 groups.

For the case where two therapies are equally superior to the other two with a hazard ratio of 0.75, then an N of 2316 (rounded up from 2315) would provide 90% power. Thus, it is conservative to power the study to detect a single isolated superior drug with HR=0.75, in which case the total sample size selected might be N= 3300 to provide 90% power.

expwlogl z was used in the following text.

However, it is also desired to conduct 6 pairwise comparisons among the 4 drug groups. Although the Hochberg closed test procedure will be employed, for the smallest nominal p-value, the adjustment is equivalent to the Bonferroni correction, i.e. a two-sided significance level of 0.05/6 being required for adjusted significance at the 0.05 level. Two group calculations with n=825 per group shows that the total N=3300 provides only 71% power to detect a HR=0.75 between any two groups with this design.

3,7 exsslogl z.sas was used in the following text.

Rather, a sample size of n=1242 per group is required to provide 90% power to detect a HR=0.75 in a two-group comparison at the 0.05/6 level under the above assumptions, thus requiring a total sample size of 4964, rounded up to N=5000 as the target enrollment. In this case, the K-1 df test of homogeneity would provide 98.3% power to detect a single superior drug group with HR=0.75, and 90% power to detect a single group with HR=0.796% .

expwlogl z was used in the following text.

It should be noted that another option might be to conduct 4 pairwise comparisons of each drug group versus the other 3 groups combined. With the smaller total N of 3300, such a test at the 0.05/4 level would provide 93% power to detect HR=0.75. However, the 6 pairwise comparisons are preferred and thus the larger sample size of N=5000 will be employed.

4 gpstratPWlogrank was used in the following text.

The study will evaluate various stratification or subgroup factors in which case a stratified-adjusted test may be conducted. To assess the effect of heterogeneity among strata, consider the case where one stratum consists of approximately 2000 subjects with a 20% lower hazard rate of 0.0875 * (0.8)=0.07/year and a smaller difference between groups with a hazard ratio of 0.85, and the other stratum consists of approximately 3000 subjects with the same risks assumed above. With the same parameters as above, and assuming that a single drug is superior to the others, then the stratum of 2000 subjects would provide an expected 122 events in the first group and 140 events in the other three groups, and the stratum of 3000 subjects would provide 199 and 251 events, respectively. The vector of stratified adjusted log hazard ratios is \beta =(-0.240713 0 0)^{T} with the first element corresponding to a hazard ratio of 0.786 for the first group versus the reference (i.e. all others). The corresponding covariance matrix V(\widehat{ {\beta}}) has diagonal elements 0.005678 for the first log odds ratio, 0.005114 for the next 2 diagonal elements, and off-diagonal elements 0.002557. The resulting non-centrality parameter is 14.58 that yields power of 90.9%. Thus, the presence of a mild group by stratum interaction (or heterogeneity) leads to some dilution of power for the 4 group test, but at an acceptable level. However, a test of no interaction or homogeneity would have a low power of only 11% to detect a difference in hazard ratios of 0.75 versus 0.85 between these two strata.

Subgroup analyses will also be performed to assess the treatment group differences between segments of the population, such as males and females, with a test of a treatment by subgroup interaction. Again assume an overall hazard rate of 0.0875 and losses at 4% per year with the first group being superior to the rest with an overall hazard ratio of 0.75 in the full cohort for one group versus the rest. Within one subgroup, assume that the hazard ratio is 25% less, i.e. a hazard ratio of 0.75 x 0.75=0.563, whereas in the other subgroup it is 25% greater, i.e. 0.75 x 1.25=0.938. For equally sized subgroups with n=2500 each, the test of homogeneity (no interaction) provides 93.9% power. For a factor with three subgroups, each with sample size 1666, the study would provide 68.9% power to detect hazard ratios of 0.563, 0.75 and 0.938.

Analyses may also be conducted involving a quantitative covariate. As for a qualitative covariate (the S strata), one analysis could assess the association of the covariate with the outcome adjusted for treatment group, and another could assess homogeneity of the covariate effect among strata (or a group by stratum interaction).

PWquantPH was used in the following text.

For N=5000, or 1250 for each of 4 groups, under the above assumptions, approximately 394 events are expected within each group (1576 total). From recent studies, such as of biomarkers in relation to cardiovascular disease, a hazard ratio of 1.4 per standard deviation change (HR_{SD}) in the covariate is desirable to detect. For a standard deviation \sigma of the covariate, the hazard ratio per unit change in the covariate is HR=HR_{SD}^{1 / \sigma} . Then the coefficient is \beta =\log (HR_{SD)} / \sigma. For \sigma =10 the Hsieh-Lavori expression with 1576 events yields virtually 100% power to detect a HR_{SD}=1.4, and 97% power to detect a smaller HR_{SD}=1.0.

4gpPWquantInt was used in the following text.

It is also desired to assess the power of the study to detect heterogeneity of a quantitative covariate effect among groups, such as with coefficient values of 1.25, 1.35, 1.45 and 1.55, the weighted average \beta =0.0333 corresponding to a % HR_{SD}=1.396. This yields a non-centrality parameter value of \psi ^{2}=10.3 on 3 for a test of homogeneity, that yields 76.7% power to detect these small differences in the hazard ratios.