Comparison of Test Procedures for Interim Analyses in Clinical Trials

Chulaluk Komoltri
Mahidol University, Clinical Epidemiology Unit, Faculty of Medicine
Siriraj Hospital, 2 Prannok Road, Bangkoknoi
Bangkok 10700, Thailand
ckomoltri@bios.unc.edu

Shrikant I. Bangdiwala, Ph.D.
University of North Carolina, Department of Biostatistics
CSCC, 137 East Franklin Street, Suite 203-B
Chapel Hill, NC 27514-4145 USA
kant@unc.edu

1. Introduction

In a randomized long-term controlled clinical trial of a new treatment, it is customary to repeatedly analyze the data for early termination of the study. To maintain type I error at the pre-specified level despite multiple hypothesis testing, various sequential approaches have been developed.

The alpha-spending function approach for group sequential designs (Lan and DeMets, 1983) allows spending the type I error at each analysis according to some pre-determined function while maintaining the overall type I error. It offers flexibility in terms of the number and timing of interim analyses. However, to determine stopping boundaries, sequential test statistics are assumed to have a multivariate normal distribution with some covariance structure.

A class of linear rank order tests (Majumdar and Sen, 1978) and unweighted and weighted tests based on the weighted empirical distribution function (EDF) (Sinha and Sen, 1982) have been developed for repeated analyses and incomplete follow-up. They have the major advantages of not requiring estimation of the covariance of sequential test statistics and not requiring specification of the number and timing of interim analyses. However, these procedures are not as well known due to the difficulty in computing their stopping boundaries.

The objective of this research is to compare four test statistics applicable for testing efficacy of an intervention in a clinical trial under interim analyses, i.e., 1) the log-rank test, 2) Gehan-Breslow's generalized Wilcoxon test, 3) linear rank test using Savage scores and standardized Wilcoxon scores, and 4) unweighted and weighted tests based on weighted EDF, all under Pocock-type and O'Brien-Fleming-type alpha-spending functions.

2. Methods

Comparison of the four test statistics was based on simulations under different scenarios and empirical application to real data, the Studies of Left Ventricular Dysfunction treatment trial. In the simulation, the study is designed as a randomized (1:1), placebo-controlled, fixed duration trial with right censored survival time as primary endpoint and 4 interim analyses at fixed calendar times. Comparisons were in terms of type I error, power and time to first reject the null hypothesis.

3. Results

Under repeated significance testing, when the two survival curves do not cross, the linear rank test with Savage scores is more powerful than with standardized Wilcoxon scores. However, due to very conservative stopping boundaries for the linear rank test, neither of the two linear rank tests is as powerful as the conventional log-rank test and the Gehan-Breslow test. Apart from that, computation of linear rank test statistics is much more time consuming.

When the two survival curves cross at some time point, use of linear rank tests with Savage and Wilcoxon scores gave about 40% and 20% respectively higher power than the log-rank test, and about 50% and 30% respectively higher power than the Gehan-Breslow test. The linear rank test with either Savage or Wilcoxon scores results in stopping the trial at the 2nd look (or 2 years in our simulated 3 year study) with treatment benefit.

4. Recommendations

In the case of no crossing between the two survival curves at any time point (even under non-proportional hazards), the log-rank test and the Gehan-Breslow test are recommended due to high power, ease of test statistic computation and critical value determination, ability to maintain type I error and insensitivity to non-differential and differential dropouts. On the other hand, with some crossing of the two survival curves, the linear rank test with Savage scores is recommended due to much higher power than the log-rank test and Gehan-Breslow test, and it is very conservative.

REFERENCES

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. Biometrika 70, 659-663.

Majumdar, H. and Sen, P.K. (1978). Nonparametric testing for simple regression under progressive censoring with staggering entry and random withdrawal. Communication in Statistics Part A - Theory and Methods 4, 349-371.

Sinha, A.N. and Sen, P.K. (1982). Tests based on empirical processes for progressive censoring schemes with staggering entry and random withdrawal. The Indian Journal of Statistics 44, 1-18.