Model-Free Tests for Hypotheses of No Quantitative and No Qualitative Interactions in Survival Analysis

Kallappa M. Koti
Food and Drug Administration, Silver Spring, MD 20993-0002

Abstract

Often it is of interest to know whether treatment effects are heterogeneous over various subgroups of patients defined by the prognostic factors. We address this problem in a survival analysis setting. We extend the standard definition of interaction used in a standard 2×2 factorial experiment to survival analysis. We define the interaction in terms of median time. We use Efron's bootstrap for estimating the standard error of median time. We propose model-free tests and discuss sample size determination. We discuss some possible applications of these tests in biomarker validation investigations.

Key Words: Stratified randomization, Kaplan-Meier method, simple effects, Gail-Simon test, likelihood ratio test, predictive biomarker.

1. Introduction

The heterogeneity of treatment effects across the levels of a baseline variable refers to the circumstance in which the treatment effects vary across the levels of the baseline characteristic. Heterogeneity is sometimes further classified as being either quantitative or qualitative. In the first case, one treatment is always better than the other, but by various degrees, whereas in the second case, one treatment is better than the other for one subgroup of patients and worse than the other for another subgroup of patients (Wang et al., 2007). Gail and Simon (1985), among others, have used the terms *crossover* interaction for qualitative interaction and *non-crossover* interaction for quantitative interaction. In this paper we propose model-free nonparametric tests for verifying the presence or absence of crossover and non-crossover interactions in survival analysis.

We focus on a clinical trial where the efficacy of an experimental treatment is compared with a control as measured by a time-to-event endpoint. We assume that the baseline characteristic is dichotomous and that randomization is stratified. This baseline characteristic is also called a stratification factor. We define the interaction in terms of median event-time. Because the distribution of survival times tends to be positively skewed, the median is the preferred summary measure of the location of the distribution. Also, the median is straightforwardly informative to the clinicians. Efron (1981) said it very nicely- "The median is often favoured as a location estimate in censored data problems because, in addition to its usual advantage of easy interpretability, it least depends upon the right tail of the Kaplan-Meier curve, which can be highly unstable if

censoring is heavy." As a result, it has become a common practice in clinical trial study reporting to give point and interval estimates for the median event-time. This motivated us to define interaction in terms of median event-time.

The tests proposed in this article may be useful in biomarker validation investigations. To facilitate a discussion, we need to state a definition of what is called a predictive biomarker. Sargent et al. (2005) define a predictive marker as a marker that predicts the differential efficacy (benefit) of a particular therapy based on marker status (e.g., only patients expressing the marker will respond to the specific treatment or will respond to a greater degree than those without the marker). That is, the treatment is better than the control in the absence and presence of the marker, but by various degrees. Therefore, quantitative interaction implies marker's predictivity. Mandrekar and Sargent (2009) state that if interactions exist, and they are fairly common in oncology clinical trials, one should plan the clinical trials in such a way that interactions can be estimated and tested. They recommend prospectively designed randomized controlled trials to test a marker-by-treatment interaction for prospective validation of a predictive marker. Mandrekar and Sargent (2009) propose two types of clinical trial designs for predictive biomarker validation when efficacy is measured by a time-to-event endpoint. One of them is a parallel group design- called the *marker by treatment interaction* design, which uses the marker status as a stratification factor when randomizing subjects to treatment. Our proposed test for quantitative interaction is applicable in the analyses of such trials.

The Gail-Simon test for qualitative interaction is widely used in practice. It has entered reference books. Marubini and Valsecchi (2004) have included a section on Gail-Simon test in their book entitled "Analyzing Survival Data from Clinical Trials and Observational Studies". Dmitrienko et al. (2005) have provided a SAS macro that calculates a p-value for the Gail-Simon likelihood ratio test for qualitative interaction. Lawrence (2003) has used the Gail-Simon test to study the inconsistency of losarton effect compared to atenolol, as measured by event-free survival (EFS), among ethnic subgroups (Black vs. Non-Black) of patients with left ventricular hypertrophy (LVH). Quan et al. (2010) have indicated a possible use of the Gail-Simon test for assessment of consistency of treatment effects in multiregional clinical trials. Our proposed test for qualitative interaction is a better alternative to the Gail-Simon test.

This article is organized as follows. In Section 2, we describe the nonparametric survival analysis setting that serves as the base for this article. In Section 2, we also explain the need for bootstrap standard error of the median and provide necessary details on its calculation. We define simple effects and interaction in terms of medians, and provide additional notations in Section 3. We propose an asymptotic *z-test* for a null hypothesis of no quantitative interaction in Section 4. In Section 5, we explain why the Gail-Simon test for no qualitative interaction fails in survival data analysis. In Section 6, we propose a new test for no qualitative interaction in terms of event-time medians. We provide a discussion on sample size determination in Section 7. We end the article with some miscellaneous comments.

2. The nonparametric preamble

We develop the tests under the frame work of a randomly right-censored survival model. We assume that $Y_1, Y_2, \cdot\cdot, Y_n$ are iid random variables with a continuous distribution

function F, and that F has a density f and median μ. These variables represent the event-times of the subjects under observation. Associated with each Y_i is an independent censoring variable C_i, which are assumed to be iid from a censoring distribution H. The data consist of n pairs (T_i, d_i), where T_i is either an observed failure-time Y_i or an observed censoring time C_i, and $d_i = I(Y_i = C_i)$. The basic quantity employed to describe time-to-event phenomenon is the survivor function $S(t) = 1 - F(t)$. The median survival time estimate is given by $m = \inf\{t: \hat{S}(t) \le 0.5\}$, where $\hat{S}(t)$ is the product-limit estimate of $S(t)$. That is, the median survival time is estimated from the product-limit estimate to be the first time that the survival curve falls to 0.5 or below. The sample median m is asymptotically normally distributed with mean μ. The variance $\sigma^2(m)$ of m is mathematically intractable. The SAS lifetest procedure provides an estimate of survivor function accompanied by survival standard error. By default, the SAS lifetest procedure uses the Kaplan-Meier method. It also produces a point estimate of the median μ of F and the 95% confidence interval- derived by Brookmeyer and Crowley (1982). Brookmeyer and Crowley obtained the confidence intervals by inverting a generalization of the sign test for censored data. They did not need the standard error of the sample median. Obviously, the SAS lifetest procedure does not provide the standard error of the sample median m. One form of the asymptotic variance of median m is

$$\sigma^2(m) = [\hat{f}(\mu)]^{-2} \times \mathrm{var}[\hat{S}(\mu)], \tag{2.1}$$

where $\hat{S}(m)$ is found using the Greenwood's formula (Collett, 1994). A slightly different version of $\sigma^2(m)$ is provided in Reid (1981):

$$\sigma^2(m) = n^{-1}\left[f(\mu)\right]^{-2}\left[\{1-F(\mu)\}^2 \int_0^{\mu} \frac{dF}{(1-F)(1-H)}\right] \tag{2.2}$$

As f is unknown, the variance $\sigma^2(m)$ given either in (2.1) or (2.2) becomes useless in estimating the population median time μ (Babu, 1985). We propose to estimate the standard error of m using the Efron's bootstrap (1981), which does not make any distributional assumptions. In a single sample setting, Efron's bootstrap may be described as follows. We draw a bootstrap sample (Y_1^*, C_1^*), (Y_2^*, C_2^*), $\cdots$, (Y_n^*, C_n^*) by independent sampling n times with replacement from F and calculate the median $m^* = m(\text{data}^*)$. We repeat this independently B times, obtaining B medians: $m^{*1}, m^{*2} \cdots, m^{*B}$. An estimated variance of the sample median time m is

$$\hat{\sigma}^2_{\mathrm{BOOT}} = \frac{1}{B-1}\left[\sum_{j=1}^{B} (m^{*j})^2 - \left(\sum_{j=1}^{B} m^{*j}\right)^2 \Big/ B \right] \tag{2.3}$$

One may set B equal to 1000. This is called "model-free" or the Efron's bootstrap procedure II. The University of Texas at Austin (1996) has provided some introductory SAS codes needed to resample a SAS dataset.

Efron (1985) states: the bootstrap estimate $\hat{\sigma}_{\text{BOOT}}$ given in (2.3) is a consistent estimate, but σ in (2.1) or in (2.2) itself may be meaningless. Therefore, we assume that $\hat{\sigma}^2_{\text{BOOT}}$, which does not depend on either f or μ is a viable substitute for $\sigma^2(m)$. Thus, we work under the notion that the sample median time m is asymptotically normally distributed with mean μ and variance $\hat{\sigma}^2_{\text{BOOT}}$. We suppress the subscript BOOT of the estimated variance.

What is an indication of an unstable median or heavy censoring is a crucial question. As observed by Brookmeyer and Crowley (1982), if the survival curve is relatively flat in the neighbourhood of 50% survival, there can be great deal of variability in the estimated median. It would be more appropriate to cite a confidence interval for the median. We propose a simple rule of thumb. If the upper limit of a 95% confidence interval on median is not available, one may conclude that median is unstable and/or censoring is heavy. Therefore, the proposed tests should work efficiently when the Brookmeyer-Crowley upper limit of a 95% confidence interval on median is available. This also minimizes the number of bootstrap samples whose Kaplan-Meier curves do not reach 0.5 survival probability. In addition, asymptotic normality requires that $m > 2\hat{\sigma}$. See also Section 8 for a related comment.

3. Basic notation, definitions, and further details

Consider a standard analysis of variance 2×2 factorial experiment with r replications. Let $\bar{y}_{00}$, $\bar{y}_{01}$, $\bar{y}_{10}$, and $\bar{y}_{11}$ denote the averages of response y, respectively, at the four treatment combinations A_0B_0, A_0B_1, A_1B_0, and A_1B_1 of two factors A and B each at 2 levels- 0 and 1. The difference $\bar{y}_{10} - \bar{y}_{00}$ represents the *simple effect* of A at B_0. Similarly, the difference $\bar{y}_{11} - \bar{y}_{01}$ is the simple effect of A at B_1. The average of simple effects is called the *main effect*. The *interaction* is defined as the difference between two simple effects (Cochran and Cox, 1957). The presence or absence of interaction is specific to the measure of the treatment effect. Interaction indicates the failure of the differences in response to changes in levels of one factor to be the same at both levels of another factor. We extend these definitions to survival analysis.

Consider a clinical trial where the experimental treatment is compared with a control as measured by a failure-time T, which is right censored. In what follows, we assume that longer failure-time implies favourable clinical outcome. We use x to denote the indicator variable for treatments. We set $x = 1$ for the experimental treatment and $x = 0$ for the control. For simplicity, we assume that there is only one stratification factor z at randomization. Let us further assume that z is dichotomous with values- 0 or 1. This defines two subsets 1 and 2 that correspond to $z = 0$ and $z = 1$, respectively. Randomization of subjects to the treatments within subsets is carried out independently.

We use the following notation.

μ_{00} : Median of F for $x = 0$ at $z = 0$
μ_{01} : Median of F for $x = 0$ at $z = 1$
μ_{10} : Median of F for $x = 1$ at $z = 0$
μ_{11} : Median of F for $x = 1$ at $z = 1$

Let $m_{ij}\,(i = 0, 1;\ j = 0, 1)$ denote the corresponding sample medians. The difference $\delta_0 = \mu_{10} - \mu_{00}$ is the simple effect of the experimental treatment at $z = 0$, and similarly, the difference $\delta_1 = \mu_{11} - \mu_{01}$ is the simple effect of the experimental treatment at $z = 1$. We define the interaction as the difference

$$\delta_1 - \delta_0 = \mu_{11} - \mu_{01} - \mu_{10} + \mu_{00} \tag{3.1}$$

The concepts of quantitative interaction and qualitative interaction are illustrated in Figures 1 and 2 in sections 4 and 6, respectively. As the parameters $\mu_{ij}\,(i = 0, 1;\ j = 0, 1)$ are unknown, the simple effects and the interaction are unknown. Estimates of the simple effects δ_0 and δ_1 are given by $D_0 = m_{10} - m_{00}$ and $D_1 = m_{11} - m_{01}$, respectively. The Kaplan-Meier estimator m_{ij} is asymptotically unbiased for μ_{ij} $(i = 0, 1;\ j = 0, 1)$. Therefore, D_0 and D_1 are unbiased for δ_0 and δ_1, respectively. The difference $D_1 - D_0$ is an unbiased estimate of the interaction between x and z. Let $\hat{\sigma}_{ij}^2$ denote the bootstrap variances of $m_{ij}\,(i = 0, 1;\ j = 0, 1)$. We use $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ to denote the bootstrap variances of D_0 and D_1, respectively. Because of the design, it follows that $\hat{\sigma}_0^2 = \hat{\sigma}_{10}^2 + \hat{\sigma}_{00}^2$ and $\hat{\sigma}_1^2 = \hat{\sigma}_{11}^2 + \hat{\sigma}_{01}^2$. Therefore, D_0 and D_1 are independently asymptotically normally distributed with means $\delta_0 = \mu_{10} - \mu_{00}$ and $\delta_1 = \mu_{11} - \mu_{01}$, and variances $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$, respectively. In what follows, z_a denotes the $100 \times a$ th percentile of the standard normal distribution.

4. Test for quantitative interaction

In order to perform the proposed test for quantitative interaction, we first calculate the four sample medians using the Kaplan-Meier method. Next, we calculate bootstrap standard errors of the medians. We skip the further computational details. Table 1 below provides the basic information needed for the test.

Table 1: Table of medians (variances) for a 2×2 factorial experiment setting

Treatment x	*Covariate z*		*Simple effect estimate*
	0	1	
0	$m_{00}\,(\hat{\sigma}_{00}^2)$	$m_{01}\,(\hat{\sigma}_{01}^2)$	$D_0 = m_{01} - m_{00}$
1	$m_{10}\,(\hat{\sigma}_{10}^2)$	$m_{11}\,(\hat{\sigma}_{11}^2)$	$D_1 = m_{11} - m_{10}$

A plot of the four points $\{(z,m):(0,m_{00}),(0,m_{10}),(1,m_{01}),(1,m_{11})\}$, in the presence of quantitative interaction should look like the one shown in Figure 1.

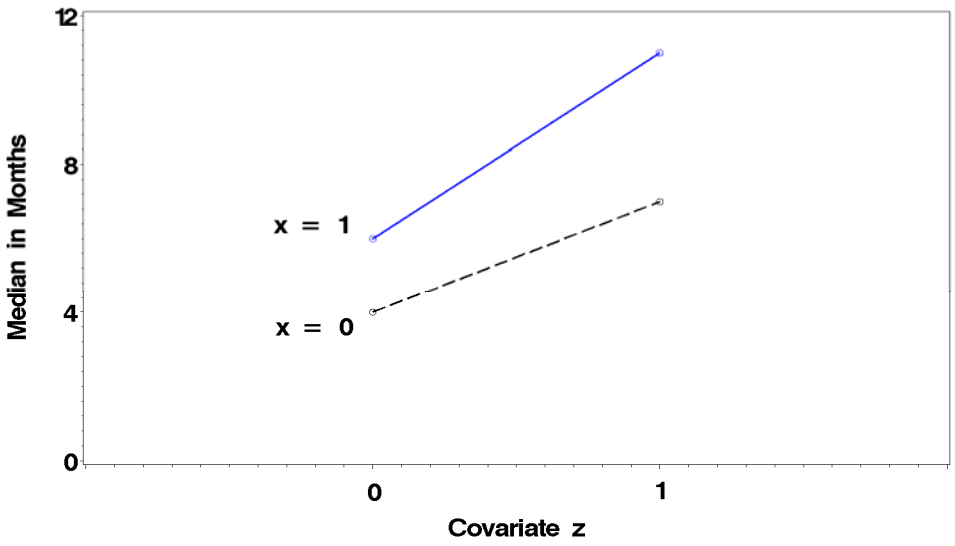

Figure 1: Illustration of quantitative interaction in a factorial arrangement

Here we want to verify if the simple effect δ_1 of the experimental treatment at $z=1$ is significantly greater than the simple effect δ_0 at $z=0$ or vice versa. As stated by Piantadosi and Gail (1993), the null hypothesis requires that a specified treatment is preferred in both subsets. We need to test the null hypothesis $H_{01}:\delta_1-\delta_0=0$ versus the two-sided alternative hypothesis $H_{A1}:\delta_1-\delta_0\neq 0$. Equivalently, we test

$$H_{01}:\mu_{11}-\mu_{01}-\mu_{10}+\mu_{00}=0 \text{ vs. } H_{A1}:\mu_{11}-\mu_{01}-\mu_{10}+\mu_{00}\neq 0 \qquad (4.1)$$

Under the null hypothesis H_{01} in (4.1), the statistic

$$D_1-D_0\equiv m_{11}-m_{01}-m_{10}+m_{00}$$

is asymptotically normally distributed with mean 0 and variance $\hat{\sigma}_0^2+\hat{\sigma}_1^2$. Therefore, we reject the null hypothesis H_{01} in favour of the alternative H_{A1} in (4.1) at level α if

$$\left| m_{11}-m_{01}-m_{10}+m_{00}\right| \Big/ \sqrt{\hat{\sigma}_0^2+\hat{\sigma}_1^2} > z_{1-\alpha/2}.$$

5. Gail-Simon test for qualitative interaction

Gail and Simon (1985) have proposed a pseudo likelihood ratio test for qualitative interaction. They have illustrated the test using a survival dataset in their section 3. Piantadosi and Gail (1993) have compared power of the Gail-Simon test with the

standardized range test proposed by Robert Tarone. They have also considered an example of survival data analysis in the power comparison. In this section we point out that Gail-Simon test that is derived in terms of *regression coefficients* is flawed and inapplicable in survival data analysis.

The Gail-Simon method is model dependent and requires that the hazard ratios in subsets 1 and 2 be estimated separately. Then in each subset i ($i = 1, 2$), the proportional hazards model sets

$$h_{Ei}(t) = h_{Ci}(t)\exp(\beta_i z), \quad t > 0, \tag{5.1}$$

where h is the hazard function and $z = 0, 1$. The model (1) can be re-written in the form $\log[h_{Ei}(t)/h_{Ci}(t)] = \beta_i z$, where the quotient $h_{Ei}(t)/h_{Ci}(t)$ is the hazard ratio (HR) of the treatment relative to the control, and β_1 and β_2 are the regression coefficients- also called the log hazard ratios. Note that the maximum likelihood estimator $\hat{\beta}_i$ of β_i ($i = 1, 2$) is asymptotically normally distributed with mean β_i and variance $\sigma_i^2 = I_i^{-1}(\beta_i)$:

$$I_i(\beta_i) = \sum_{j=1}^{k} \frac{r_{j1} r_{j2} \exp(\beta_i)}{[r_{j1}\exp(\beta_i) + r_{j2}]^2}, \tag{5.2}$$

where r_{jE} and r_{jC} denote the numbers of subjects at risk in the experimental treatment and control at the smallest failure time $t_{(j)}$ ($j = 1, \cdots, k$), respectively (Collett, 1994).

The objective is to test the hypothesis that there is no crossover interaction. Gail and Simon used $\Delta = \{\boldsymbol{\beta} : -\infty < \beta_1, \beta_2 < \infty\}$ to denote the unrestricted parameter space. They set the null parameter space as $\Delta' = O^+ \cup O^-$, where $O^+ = \{\boldsymbol{\beta} : \beta_1 \geq 0, \beta_2 \geq 0\}$, and $O^- = \{\boldsymbol{\beta} : \beta_1 \leq 0, \beta_2 \leq 0\}$. Gail and Simon consider the pseudo likelihood

$$\ell(\boldsymbol{\beta}) = c_1(\boldsymbol{\beta}) \times \exp \sum_{i=1}^{2} [-(\hat{\beta}_i - \beta_i)^2 / 2\sigma_i^2], \tag{5.3}$$

where $c_1(\boldsymbol{\beta}) = (2\pi\sigma_1\sigma_2)^{-1}$, for deriving their test. As seen from (5.2), σ_i^2 is a function of β_i ($i = 1, 2$), and therefore, c_1 depends on $\boldsymbol{\beta}$.

Gail-Simon proposed the likelihood ratio test statistic

$$\lambda_{GS} = \frac{\max_{\boldsymbol{\beta} \in \Delta'} [\exp \sum_{i=1}^{2} (-(\hat{\beta}_i - \beta_i)^2 / 2\sigma_i^2)]}{\max_{\boldsymbol{\beta} \in \Delta} [\exp \sum_{i=1}^{2} (-(\hat{\beta}_i - \beta_i)^2 / 2\sigma_i^2)]} \tag{5.4}$$

They state that the hypothesis of no crossover interaction is rejected if both

$$Q^- = \sum (\hat{\beta}_i^2 / \sigma_i^2) I(\hat{\beta}_i > 0) > c \text{ and } Q^+ = \sum (\hat{\beta}_i^2 / \sigma_i^2) I(\hat{\beta}_i < 0) > c,$$

where $I(\hat{\beta}_i > 0) = 1$ if $\hat{\beta}_i > 0$ and 0 otherwise, and $I(\hat{\beta}_i < 0) = 1$ if $\hat{\beta}_i < 0$ and 0 otherwise. The quantities Q^+ and Q^- are the minimum values of $\Sigma(\hat{\beta}_i - \beta_i)^2 / \sigma_i^2$ over O^+ and O^-, respectively, and the likelihood ratio test can be expressed as $\min(Q^+, Q^-) > c$. They provide the values of c corresponding to significance levels $0.001,\ 0.05,\ 0.1,$ and 0.2. See Gail and Simon (1985) for further details.

We have the following comments on their test. The intended likelihood ratio for the Gail-Simon test is

$$\lambda = \frac{\max_{\boldsymbol{\beta} \in \Delta'} [\ell(\boldsymbol{\beta}); \boldsymbol{\beta} \in \Delta']}{\max_{\boldsymbol{\beta} \in \Delta} [\ell(\boldsymbol{\beta}); \boldsymbol{\beta} \in \Delta]}, \tag{5.5}$$

where $\ell(\boldsymbol{\beta})$ is given by (5.3). This statistic λ in (5.5) is different from λ_{GS} in (5.4) in that $c_1(\boldsymbol{\beta}) = (2\pi\sigma_1\sigma_2)^{-1}$ is missing from the numerator and from the denominator of λ_{GS}. The *restricted* maximum likelihood estimate is not equal to the *unrestricted* maximum likelihood estimate. Therefore, the c_1 in the numerator and c_1 in the denominator of the statistic λ in (5.5) cannot be cancelled. Also the variance σ_i^2 in the exponent in $\ell(\boldsymbol{\beta})$ cannot be ignored. Gail and Simon have assumed that the variances σ_i^2 are known. At the end of section 2 they have claimed that consistent estimates s_i^2 may be inserted for σ_i^2 in all the previous equations without altering the asymptotic distribution theory. In the current scenario σ_i^2 is not a *nuisance* parameter. As seen from (5.2), σ_i^2 is an explicit function of β_i on which inference is needed. If β_i is unknown, σ_i^2 is unknown. A rigorous derivation of the test requires that σ_i^2 be treated as a function of β_i. In their Table 3, $s_i = 0.208$ is the standard error of $\hat{\beta}_i$ for the stratum "Age < 50 and PR <10". An important question is: how did they get it? In practice, two different values of s_i are used to make inference on β_i (SAS Institute, Inc., 1997). A numerical value of s_1, for example, may be obtained using the observed information in (5.2) evaluated at $\beta_1 = 0$. It is because for testing, $H_0 : \beta_1 = 0$, the score test uses the observed information in (5.2) evaluated at $\beta_1 = 0$. Setting the β s in c_1 as 0 and then maximizing the numerator and the denominator of λ_{GS} would be erroneous. Alternatively, for the Wald test, the standard error s_1 is calculated using the observed information in (5.2) evaluated at $\beta_1 = \hat{\beta}_1$. An attempt to estimate σ_1^2 as the inverse of $I(\hat{\beta}_1)$ and then treat it as a constant in the derivation of the test is not a good move. The denominator in (5.5) is not equal to 1 as claimed by Gail and Simon. In addition, the Gail-Simon critical values c are valid only if the conditional distribution of $\hat{\beta}_i^2 / s_i^2$ given $\hat{\beta}_i < 0$ or $\hat{\beta}_i > 0$ is central chi-square with 1 degree of freedom. As s_i^2 is a function of $\hat{\beta}_i$, $\hat{\beta}_i$ and s_i^2 are not independent ($i = 1, 2$), and therefore, conditional distribution of $\hat{\beta}_i^2 / s_i^2$ is not central chi-square. Consequently, derivation of the Gail-Simon test fails when applied in survival analysis. In the next section we modify the Gail-Simon test and propose a new test for the null hypothesis of no crossover interaction in terms of medians instead of regression coefficients β_1 and β_2.

6. New test for qualitative interaction

Once again, the basic information needed is the same as shown in Table 1. But this time a plot of the four points $\{(z,m):(0,m_{00}),(0,m_{10}),(1,m_{01}),(1,m_{11})\}$, in the presence of qualitative interaction should look like the one shown in Figure 2.

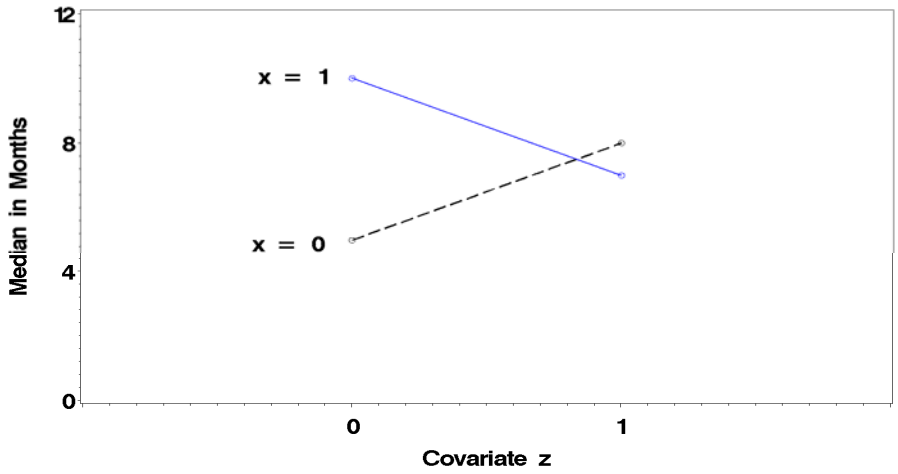

Figure 2: Illustration of qualitative interaction in a factorial arrangement

We test the null hypothesis $H_{02}: \mu_{10} \geq \mu_{00}$ and $\mu_{11} \geq \mu_{01}$ or $\mu_{10} \leq \mu_{00}$ and $\mu_{11} \leq \mu_{01}$ vs. the alternative hypothesis H_{A2} that H_{02} is false. Here the alternative hypothesis H_{A2} means: $\mu_{10} \geq \mu_{00}$ and $\mu_{11} \leq \mu_{01}$ and vice versa. In Piantadosi and Gail's (1993) words, the null hypothesis requires that the treatment effect has the same unspecified direction in both subsets. In order to derive our test, we switch back to the Gail-Simon original notation O^+ and O^-. Let $\boldsymbol{\delta} = (\delta_0, \delta_1)^T$, where δ_0 and δ_1 are the simple effects, which were defined in earlier in Section 3. Let $O^+ = \{\boldsymbol{\delta}: \delta_0 \geq 0, \delta_1 \geq 0\}$ and $O^- = \{\boldsymbol{\delta}: \delta_0 \leq 0, \delta_1 \leq 0\}$. We let $\Delta' = O^+ \cup O^-$ denote the parameter space restricted under the null hypothesis and Δ denote the unrestricted parameter space. The likelihood ratio test statistic is

$$\lambda_{LR} = \frac{\max_{\boldsymbol{\delta}\in\Delta'}(2\pi\hat{\sigma}_1\hat{\sigma}_2)^{-1}\,[\exp\sum_{i=0}^{2}(-(D_i-\delta_i)^2/2\hat{\sigma}_i^2)]}{\max_{\boldsymbol{\delta}\in\Delta}(2\pi\hat{\sigma}_1\hat{\sigma}_2)^{-1}\,[\exp\sum_{i=0}^{2}(-(D_i-\delta_i)^2/2\hat{\sigma}_i^2)]} \qquad (6.1)$$

As the bootstrap variances $\hat{\sigma}_i^2$ of D_i are independent of δ_i $(i = 0, 1)$, the factor $(2\pi\hat{\sigma}_1\hat{\sigma}_2)^{-1}$, which appears in the numerator and in the denominator is cancelled out. The likelihood ratio (LR) test statistic in (6.1) can now be written as

$$\lambda_{LR} = \frac{\max_{\boldsymbol{\delta}\in\Delta'}[\exp\sum_{i=0}^{2}(-(D_i-\delta_i)^2/2\hat{\sigma}_i^2)]}{\max_{\boldsymbol{\delta}\in\Delta}[\exp\sum_{i=0}^{2}(-(D_i-\delta_i)^2/2\hat{\sigma}_i^2)]} \tag{6.2}$$

This test statistic is similar to the Gail-Simon test statistic λ_{GS} in (5.4). Therefore, Gail-Simon arguments used in support of their test statistic λ_{GS} prove to be useful here in deriving the test λ_{LR} given in (6.2). We reject the hypothesis of no crossover interaction if both

$$Q^- = \sum_i (D_i^2/\hat{\sigma}_i^2)\times I(D_i>0) > c \text{ and } Q^+ = \sum_i (D_i^2/\hat{\sigma}_i^2)\times J(D_i<0) > c\,,$$

where $I(D_i>0)=1$ if $D_i>0$ and 0 otherwise, and $J(D_i<0)=1$ if $D_i<0$ and 0 otherwise. Gail-Simon (1985) critical values (c) remain valid for our new test as well.

7. Sample size for quantitative interaction test

Testing for interaction arises in clinical trials where comparing a time-to-event between two treatments is a primary objective. Consider such a study where the subjects are randomized to the treatments in a 1:1 ratio. The required number of deaths d can be obtained from the equation: $d = 4(z_{\alpha/2}+z_\beta)^2/(\log \mathrm{HR})^2$, where HR is the hazard ratio of the active treatment relative to the control. See p. 255 in Collett (1994) for details.

Mandrekar and Sargent (2009) have discussed sample size consideration for a marker by treatment interaction design. They have calculated the sample size as the sum of the number of events required for the marker +ve arm comparison and the number of events required for the marker −ve arm comparison. That is,

$$d = 4(z_{\alpha/2}+z_\beta)^2 \times \left[(\log\frac{\mu_{10}}{\mu_{00}})^{-2} + (\log\frac{\mu_{11}}{\mu_{01}})^{-2}\right].$$

We base the sample size determination on the interaction. We assume that failure-time T has an exponential distribution. That is, $f(t)=\lambda\exp(-\lambda t);\ t>0,\ \lambda>0$. The maximum likelihood estimator of λ is $\hat{\lambda} = d/\sum_1^n t_r$, where d is the number of events out of the n observations. The median estimate is $\log 2/\hat{\lambda}$. The maximum likelihood estimator $\hat{\lambda}$ is asymptotically normally distributed. By the delta method, the sample median $\log 2/\hat{\lambda}$ is approximately normally distributed with mean $\log 2/\lambda$ and standard error equal to $\log 2/\lambda\sqrt{d}$ (Collett, 1994). Let λ_{ij} denote the parameters of the exponential distributions for the failure-times corresponding to subsets formed by $x=i,\ i=0,1$, and $z=j,\ j=0,1$. We assume that the medians for the two treatment arms at $z=0$ are known. That is, we assume that λ_{00} and λ_{10} are pre-specified and therefore, $\delta_0 = (1/\lambda_{10} - 1/\lambda_{00})\log 2$ is known. Let γ be a pre-specified constant that represents the interaction of interest. Set $\delta_1 = \delta_0 + \gamma$. Our objective is to test $H_0 : \delta_1 - \delta_0 = 0$ versus

$H_A : \delta_1 - \delta_0 = \gamma$ at α level of significance. We seek a power of $1-\beta$. For simplicity, we assume that both subsets have the same number of events d per arm so that the total number of events needed is at least $4d$. It follows that under H_0, the variance of $D_1 - D_0$ is $\sigma_0^2 = 2(\log 2)^2(\lambda_{10}^{-2} + \lambda_{00}^{-2})/d$. Clearly, we are in the single sample z-test setting. Therefore, $d = [\sigma_0(z_{\alpha/2} + z_\beta)/|\gamma|]^2$. The test for interaction should be performed at a higher significance level, preferably between 0.1 and 0.2; and the power may not exceed 0.8.

8. Miscellaneous comments

Often an estimate of median is available but the upper limit of the 95% confidence is not available. In this case, median may not be available for a good number of bootstrap samples that would be used to find standard error of the sample median. For these samples, the median estimate may be set to be the largest failure time (Keaney and Wei, 1994). In this case, one should check that $m > 2\sigma$ for both treatments in each subset.

Demonstration of overall treatment effect in the target population is the ultimate goal of a clinical trial. Stratified log-rank test is the most commonly used to compare survival curves. Stratified Cox model based hazard ratio, which requires a proportional hazards assumption, is often used as a measure of clinical benefit of the experimental treatment relative to the control. The concept of hazard ratio is elusive. Clinicians find it hard to understand. In view of this, we define the term of what is called the *main effect*. The average $(\delta_0 + \delta_1)/2$, of the simple effects is called the main effect of the experimental treatment. The mean $(D_0 + D_1)/2$ is an asymptotically unbiased estimate of the main effect. The main effect estimate is model-free and is more informative and appealing compared to a hazard ratio. The estimated main effect may be numerically different from the difference between the treatment medians, which are routinely obtained from the combined data.

Acknowledgements

This article reflects the views of the author and should not be construed to represent FDA's views or policies. No official support or endorsement of this article by the Food and Drug Administration is intended or should be inferred. The author is thankful to Dr. Rajeshwari Sridhara, FDA/CDER/DBV, for the support

References

Babu, G. J. (1985). A note on bootstrapping the variance of sample quantiles. *Annals of the Institute of Statistical Mathematics.* 38 (A): 439-443.

Brookmeyer, R. and Crowley, J. (1982). A confidence interval for the median survival time. *Biometrics*, 38, 29-41.

Cochran, W. G. and Cox, G. M. (1957). *Experimental designs*. John Wiley & Sons: New York.

Collett, D. (1994). *Modelling survival data in medical research.* Chapman & Hall: New York, 1994.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association,* 76, 312-319.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide.* SAS Institute, Inc. Cary, NC.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics,* 41: 361-372.

Keaney, K. M. and Wei, L. J. (1994). Interim analyses based on median survival times. *Biometrika*, 81 (2): 279-286.

Lawrence, J. (2003). Analysis of LIFE Study by ethnic demographic subgroup. http://www.fda.gov/ohrms/dockets/ac/03/slides/3920s1.htm.

Mandrekar, S. J. and Sargent, D. J. (2009). Clinical trail designs for predictive biomarker validation: one size does not fit all. *Journal of Biopharmaceutical Statistics*, 19: 530-542.

Marubini, E. and Valsecchi, M. G. (2004). *Analyzing survival data from clinical trials and observations.* John Wiley & Sons. West Sussex, England.

Piantadosi, S. and Gail, M. H. (1993). A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12: 1239-1248.

Quan, H., Li, M., et al. (2010). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal*, 44: 617-632.

Reid, N. (1981). Estimating the median survival time. *Biometrika*, 68: 601-608.

Sargent, D. J., Conley, B. A., Allegra, C. and Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of clinical oncology*, 23: 2020-2027.

SAS Institute Inc. (1997). *SAS/STAT Software: Changes and Enhancements through Release 6.12.* SAS Institute Inc., Cary, NC, USA.

The University of Texas at Austin. (1996). Setting and resampling in SAS. http://ftp.sas.com/techsup/download/stat/jackboot.htm/

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. and Drazen, J. M. (2007). Statistics in medicine – reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, 357 (21): 2189-2194.