

On P-Value Calculation for Multi-Stage Additive Tests

Jun Sheng

*School of Statistics
University of Minnesota
224 Church Street SE
Minneapolis, MN 55455*

Peihua Qiu

*School of Statistics
University of Minnesota
224 Church Street SE
Minneapolis, MN 55455*

June 19, 2006

Abstract

Multi-stage additive tests are commonly used in applications. Appropriate definition of its test decisions, however, turns out to be challenging. There are a number of existing methods for this purpose, mainly by combining p -values of individual tests using a conditional error probability function. While these methods are flexible enough to use in most applications, their results depend on the conditional error probability function and selection of this function is often subjective. Motivated by a research problem regarding comparison of two survival hazard rate functions, in this paper, we suggest using an alternative and simpler definition of the overall p -value of a multi-stage additive testing procedure. By this definition, no conditional error probability function needs to be chosen, the entire testing procedure is easy to interpret, and the overall p -value has an explicit formula for a general multi-stage additive testing procedure.

Key Words: Combination test; Crossing hazard rates; Group sequential test; Overall p -value; Significance level; Survival Analysis; Two-stage test.

1 Introduction

Multi-stage additive tests become popular in applications, especially in the field of clinical trials, for achieving maximal flexibility in trial conduct and minimal patient exposure and economic expenditure (e.g., Pocock 1977, O'Brien and Fleming 1979, DeMets and Ware 1980, 1982, Wang and Tsiatis 1987, Pampallona and Tsiatis 1994). Appropriate definition of their decision rules and overall p -values, however, turns out to be challenging (e.g., Cui et al. 1999, Lehmacher and Wassmer 1997, Brannath et al. 2002). This paper focuses on this problem.

In the literature, there are some existing methods for defining decision rules and overall p -values of multi-stage additive tests, by combining p -values of tests in individual stages (e.g., Bauer and Köhne 1994), or by specifying a conditional error probability function (e.g., Proschan and Hunsberger 1995). While these methods are flexible enough to use in most applications, their results depend on the conditional error probability function or the function for combining the p -values of individual tests. Selection of such functions and the parameters involved is often subjective. Due to the relatively complicated structure of these methods, the parameters involved may not be easy to interpret, and sometimes the overall p -values are not convenient to compute when the number of stages is large. More introduction about existing methods can be found in Section 2.2.

In this paper, we suggest using an alternative and simpler definition of the overall p -value for a multi-stage additive testing procedure in cases when the individual tests are independent or when they have the property of p -clud (cf., Section 2.1 for introduction). This definition is based mainly on properties of conditional probabilities when multiple events are involved. By this method, the overall p -value of a multi-stage additive testing procedure has an explicit formula; thus, it is convenient to compute. That formula does not depend on any extra parameters, besides the significance levels of tests in individual stages, which are also used in most existing methods.

The proposed method is described in detail in Section 2. Connections and differences between this method and some existing methods are also explained there. Then, it is demonstrated in the example to compare two hazard rates of survival data in Section 3. Finally, some remarks conclude the article in Section 4.

2 Significance Level and p -Value of A Multi-Stage Additive Test

This section is organized in two parts. The proposed method for defining decision rules of a multi-stage additive test is introduced in Section 2.1, and its connection to some existing methods is described in Section 2.2.

2.1 The proposed method

Suppose that the significance level of a k -stage procedure is α , for an integer $k \geq 2$, and the significance levels of the tests in the individual stages are $\alpha_1, \alpha_2, \dots, \alpha_k$, respectively. Then, the k -stage procedure rejects H_0 if and only if the test in the first stage rejects H_0 , or it fails to reject H_0 but the test in the second stage rejects H_0 , or both the first two tests fail to reject H_0 but the test in the third stage rejects H_0 , and so forth. If the tests in the individual stages are independent of each other, then by the definition of these significance levels and by recursively using the properties of conditional probabilities, we have

$$\sum_{j=1}^k \alpha_j \prod_{\ell=1}^{j-1} (1 - \alpha_\ell) = \alpha, \quad (2.1)$$

where $\alpha_0 = 0$. Therefore, as long as tests in the individual stages of a k -stage procedure are independent of each other and the significance levels of the individual tests satisfy equation (2.1), the overall significance level of the k -stage procedure would be controlled at α .

To define the overall p -value of the k -stage procedure, let p_1, p_2, \dots, p_k be the p -values of the tests in the k individual stages. Then we define the following function $g(p_1, p_2, \dots, p_k)$ to be the overall p -value:

$$g(p_1, p_2, \dots, p_k) = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1 \\ \alpha_1 + p_2(1 - \alpha_1), & \text{if } p_1 > \alpha_1 \text{ and } p_2 \leq \alpha_2 \\ \vdots & \vdots \\ \sum_{j=1}^{k-1} \alpha_j \prod_{\ell=1}^{j-1} (1 - \alpha_\ell) + p_k \prod_{\ell=1}^{k-1} (1 - \alpha_\ell), & \text{if } p_1 > \alpha_1, \dots, p_{k-1} > \alpha_{k-1}, \end{cases} \quad (2.2)$$

where $\alpha_1, \alpha_2, \dots, \alpha_k \in [0, 1]$ are parameters.

From (2.2), it can be seen that, if the k individual stages of the k -stage procedure are independent of each other, $\alpha_1, \alpha_2, \dots, \alpha_k$ are their significance levels as specified in (2.1), then the test

defined by (2.2) is equivalent to the one defined by (2.1). More specifically, in such cases, the test defined by (2.2) has size α if $\alpha_1, \alpha_2, \dots, \alpha_k$ satisfy (2.1), its overall p -value $g(p_1, p_2, \dots, p_k)$ has uniform null distribution on $[0, 1]$ if the p -values p_1, p_2, \dots, p_k all have uniform null distributions on $[0, 1]$, and the parameters $\alpha_1, \alpha_2, \dots, \alpha_k$ are the significance levels of the individual tests. This result can be generalized to cases when the p -values p_1, p_2, \dots, p_k have the property of *p-clud* (cf., e.g., Brannath et al. 2002) that, for any constant $c \in [0, 1]$,

$$\begin{aligned} P_{H_0}(p_1 \leq c) &\leq c, \\ P_{H_0}(p_j \leq c | p_1, p_2, \dots, p_{j-1}) &\leq c, \text{ for } j = 2, 3, \dots, k. \end{aligned} \quad (2.3)$$

The *p-clud* property (2.3) implies that the null distribution of p_1 and the null conditional distributions of p_j given $(p_1, p_2, \dots, p_{j-1})$ are all not smaller than the Uniform distribution on $[0, 1]$. Under this assumption and the assumption that the parameters $\alpha_1, \alpha_2, \dots, \alpha_k$ satisfy condition (2.1), it can be checked that the test defined by (2.2) has size α . Namely,

$$P_{H_0}(g(p_1, p_2, \dots, p_k) \leq \alpha) \leq \alpha.$$

In equations (2.1) and (2.2), the parameters $\alpha_1, \alpha_2, \dots, \alpha_k$ still need to choose when the proposed method is used in an application. If we have some prior information about the sizes of the individual tests, then such information should be used in the selection of these parameters (cf., the example discussed in Section 3). If we do not have such prior information, then we can simply let $\alpha_1 = \alpha_2 = \dots = \alpha_k$. For instance, when $k = 2$, α_1 and α_2 can be chosen to be

$$\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha}. \quad (2.4)$$

2.2 Connection to some existing methods

Let us focus on two-stage tests for simplicity. Following the notation of Brannath *et al.* (2002), let α_0 and α_1 be two predetermined decision boundaries, satisfying $0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$, where α is the size of the two-stage test. Then, most existing methods define the decision rules in the following way. In the first stage, we reject H_0 if $p_1 \leq \alpha_1$, and fail to reject H_0 if $p_1 > \alpha_0$. In both cases, the whole testing procedure stops. Otherwise, if $\alpha_1 < p_1 \leq \alpha_0$, the second-stage proceeds as follows. Let $C(p_1, p_2)$ be a predetermined function on space $(p_1, p_2) \in (\alpha_1, \alpha_0] \times [0, 1]$, which has the properties that (i) $C(p_1, p_2)$ increases with both arguments, (ii) it is a strictly increasing

function in at least one argument, and (iii) it is left continuous in p_2 . Then, in the second stage, H_0 is rejected if and only if $C(p_1, p_2) \leq c$, where c is determined by

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \leq c\}} dy dx = \alpha, \quad (2.5)$$

and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The overall p -value of the combination test (cf., Tsiatis *et al.* 1984) is often defined by

$$q(p_1, p_2) = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0, \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \leq C(p_1, p_2)\}} dy dx, & \text{otherwise.} \end{cases} \quad (2.6)$$

In the literature, there are several existing methods having the framework described above. One well known method of this type is the Fisher's weighted product test (Fisher 1932), which corresponds to the function

$$C(p_1, p_2) = p_1^w \cdot p_2, \quad (2.7)$$

where $w > 0$ is an unknown parameter. Then, the constant c defined in equation (2.5) becomes

$$c = \begin{cases} \frac{\alpha - \alpha_1}{\ln \alpha_0 - \ln \alpha_1}, & \text{if } w = 1, \\ \frac{(\alpha - \alpha_1)(1-w)}{\alpha_0^{1-w} - \alpha_1^{1-w}}, & \text{otherwise.} \end{cases} \quad (2.8)$$

When $\alpha_0 = 1$ and $w = 1$, the corresponding overall p -value of this test has the form

$$q(p_1, p_2) = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1, \\ \alpha_1 - p_1 p_2 \cdot \ln \alpha_1, & \text{if } p_1 > \alpha_1 \text{ and } p_1 p_2 \leq \alpha_1, \\ p_1 p_2 - p_1 p_2 \cdot \ln(p_1 p_2), & \text{if } p_1 > \alpha_1 \text{ and } p_1 p_2 > \alpha_1. \end{cases} \quad (2.9)$$

Another example is the weighted inverse normal method by Lehman and Wassmer (1999), which corresponds to the function

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)], \quad (2.10)$$

where w_1 and w_2 are two positive weights satisfying $w_1^2 + w_2^2 = 1$, Φ is the cumulative distribution function of the standard normal distribution, and Φ^{-1} is its inverse function.

There are some other definitions of $C(p_1, p_2)$ (cf., e.g., Proschan and Hunsberger 1995). A common feature of these existing methods is that they depend on some parameters, e.g., w in equation (2.7) and w_1 and w_2 in equation (2.10). These parameters usually do not have intuitive and simple explanations. These methods can be generalized to cases of general multi-stage procedures.

But computation of the overall p -value in such cases is often difficult. For instance, the integration in (2.6) may not be easy to compute if the method (2.10) is used for a k -stage procedure with k large. Similarly, formulas (2.8) and (2.9) would become quite complicated in such cases.

As a comparison, for two-stage procedures, the proposed method corresponds to $C(p_1, p_2) = p_2$ with $\alpha_0 = 1$, which is not included in (2.7) and (2.10) since w , w_1 , and w_2 there can not be 0. Using the proposed method, no extra parameters are needed, besides α_1 and α_2 which are significance levels of individual stages and which are used in most existing methods (cf., equation (2.6)). When the proposed method is used for general multi-stage procedures, its overall p -value has an explicit formula (cf., equation (2.2)). Therefore, it is convenient to compute.

When $k = 2$, $\alpha = 0.05$, $\alpha_0 = 1$, and α_1 and α_2 are determined by (2.4) to be both 0.0253, Figures 2.1(a)–2.1(c) shows the overall p -values of the Fisher’s weighted product test (cf., formulas (2.7)–(2.9)), when $w = 0.1, 1$, and 10 , respectively. It can be seen that the shape of the overall p -value surface depends on w . Similarly, the overall p -value surfaces by Lehmacher and Wassmer’s method (cf., formula (2.10)) are shown in Figures 2.1(d) and 2.1(e) when $w_1 = 0.5$, and 0.9 , respectively. Their shapes also depend on the parameter w_1 . The overall p -value surface of the proposed method is shown in Figure 2.1(f).

3 Application to Comparison of Two Hazard Rates

In this section, we apply the proposed method to the problem of comparing two hazard rates. Let h_1 and h_2 be the hazard rates of the survival times of subjects in the control and treatment groups, respectively, and let $[0, \tau]$ be the time range of interest. Then, we are interested in testing

$$\begin{aligned}
 &H_0 : h_1(t) = h_2(t), \text{ for all } t \in [0, \tau] \\
 \text{vs. } &H_a : h_1 \text{ and } h_2 \text{ are different on } [0, \tau].
 \end{aligned} \tag{3.1}$$

It is well known that conventional testing procedures, such as logrank, Gehan-Wilcoxon, and Peto-Peto procedures (cf., e.g., Klein and Moeschberger 1997, Chapter 7), are powerful for this purpose only when the hazard rates do not cross; some existing procedures designed for handling the crossing hazard rates problem, including those by Anderson and Senthilselvan (1982), Breslow *et al.* (1984), Lin and Wang (2004), and Liu *et al.* (2006), are powerful only when the hazard rates cross. To test the general alternative in (3.1), recently Qiu and Sheng (2005) suggested a two-stage additive

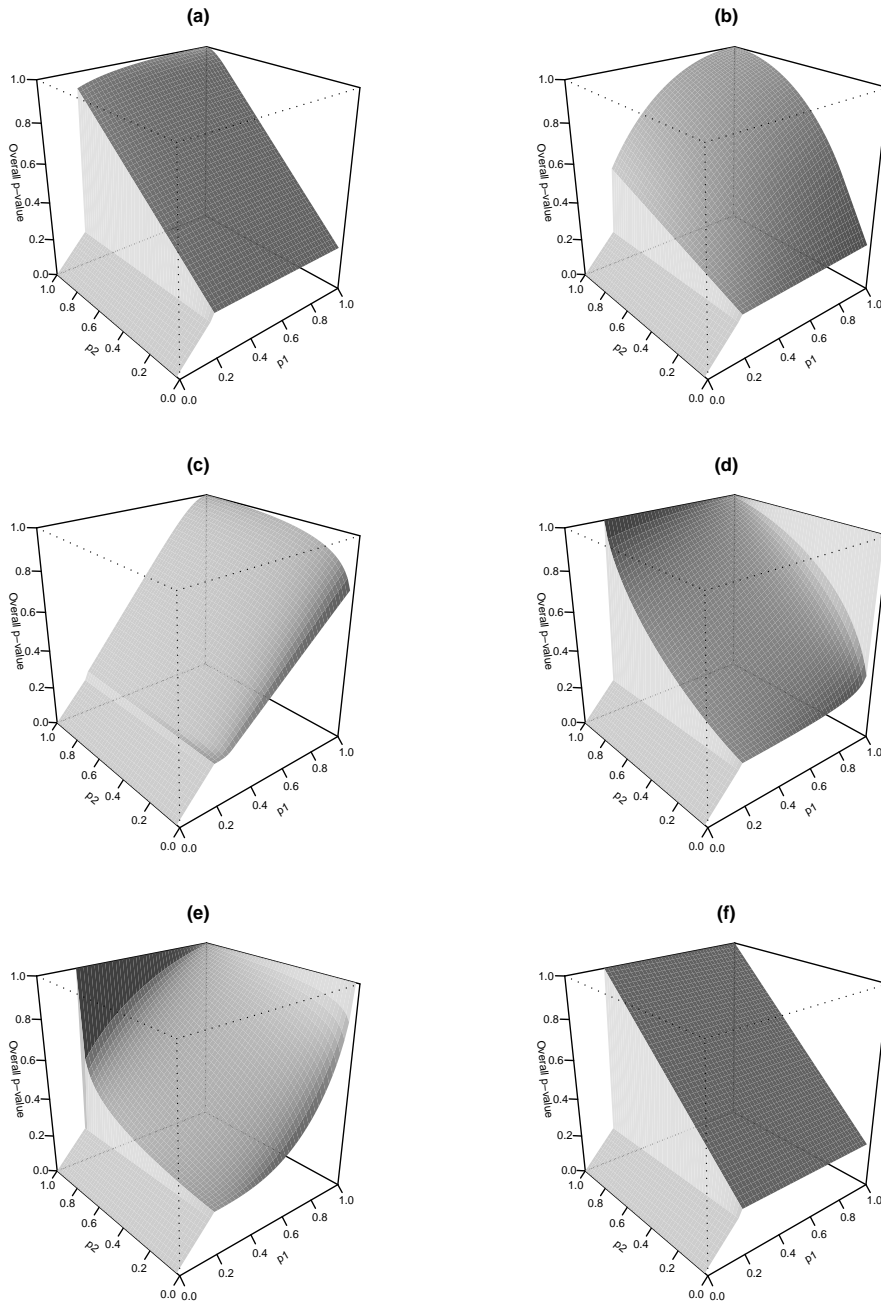


Figure 2.1: (a)-(c): Overall p-value surfaces of Fisher's weighted product test when $w = 0.1, 1,$ and $10,$ respectively. (d)-(e): Overall p-value surfaces of Lehmacher and Wassmer's method when $w_1 = 0.5,$ and $0.9,$ respectively. (f): Overall p-value surface of the proposed method. For all methods, α_0 is set to be $1.$

procedure. In the first stage, the conventional logrank procedure is applied to detect any non-crossing differences between h_1 and h_2 . If we fail to reject H_0 in the first stage, then, in the second stage, a test designed for detecting crossing differences between h_1 and h_2 is applied, which is proved to be independent of the conventional logrank test.

In this example, if we have some prior information about the pattern of the two hazard rates, then α_1 can be determined accordingly. In the two extreme cases that we believe based on our past experience the two hazard rates can not cross each other or they can not run parallel to each other, α_1 can be simply chosen α and 0, respectively. If we do not have any such information, then α_1 and α_2 can be chosen the same, as specified in (2.4).

Next, we apply this two-stage procedure to two real datasets, using the proposed method and the existing methods (2.7) and (2.10), respectively, to determine its decision rules. The first dataset is about kidney dialysis patients, which was taken from a study designed to assess the time to the first exit-site infection (in months) in 119 patients with renal insufficiency. Among all patients, 43 of them utilized a surgically placed catheter (Group 1) and 76 of them utilized a percutaneous placement of their catheter (Group 2). Catheter failure was the primary reason for censoring. There were 27 censored observations in Group 1 and 65 in Group 2. This dataset is described in detail by Klein and Moeschberger (1997, section 1.4). The second dataset is obtained from a study about the tumorigenesis of a drug reported by Mantel et al. (1977). In the experiment, rats were taken from fifty distinct litters and of each litter, one rat was randomly selected and given the drug, another two rats were selected as controls and were given a placebo. All mice are females. The number of censored observations are 29 in the treatment group and 81 in the control group. The life-table estimators of the hazard rates in these two cases are shown in Figures 3.1(a) and 3.1(b), respectively, from which it can be seen that they cross each other in the first case and run parallel to each other in the second case.

When using the two-stage procedure, we let $\alpha = 0.05$, and $\alpha_1 = \alpha_2 = 0.0253$, by equation (2.4). The p -values p_1 and p_2 of the tests in the two individual stages, and the overall p -values computed by the proposed method and the existing methods (2.7) and (2.10) using several different parameters are shown in Table 3.1. From that table, it can be seen that, for the kidney dialysis patients data, the first individual test is not significant and the second individual test is significant. The overall p -value computed by the proposed method is smaller than α ; but those computed by (2.7) and (2.10) depend on their parameter values. When w and w_1 are small, their overall p -values

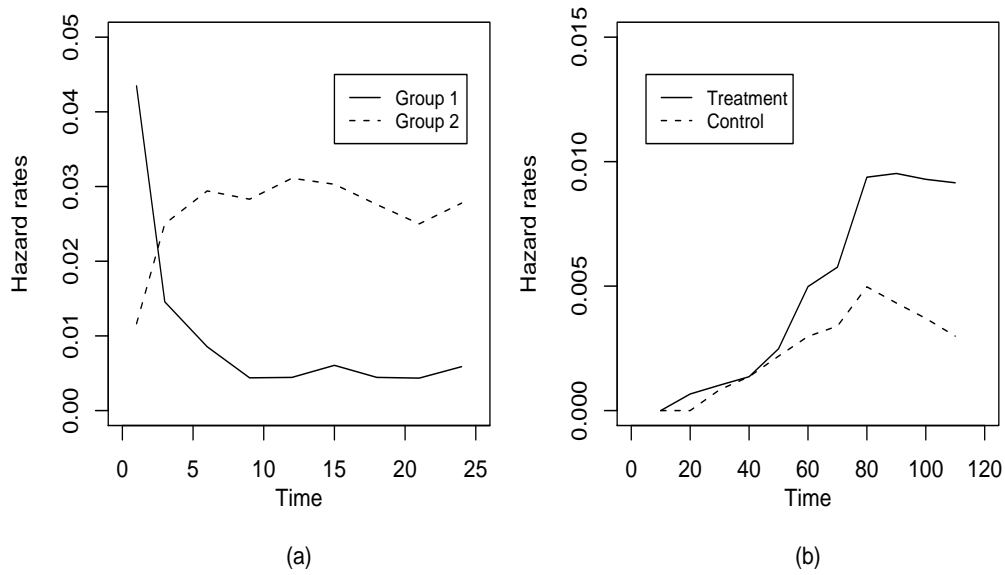


Figure 3.1: (a) Life-table estimators of the hazard rates for the kidney dialysis patients data. (b) Life-table estimators of the hazard rates for the rat data.

are small than α . In cases when w and w_1 are large, their overall p -values are larger than α . For the rats data, the first individual test is significant. In such cases, the overall p -values computed by various methods are all the same.

Table 3.1: P-values of various procedures when they are applied to the kidney dialysis patients data (Kidney) and the rats data (Rat). The number 0.106 is put in parenthesis to denote the case when that individual p -value does not need to compute since the whole test stops at stage one.

Dataset	Stage-I	Stage-II	Fisher			L-W			New
			$w = 0.1$	$w = 1$	$w = 10$	$w_1 = 0.5$	$w_1 = 0.9$	$w_1 = 0.99$	
Kidney	0.1120	0.0010	0.0262	0.0257	0.0624	0.0256	0.0264	0.0506	0.0263
Rat	0.0030	(0.1060)	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030

4 Concluding Remarks

We have presented a definition of the decision rules for multi-stage additive testing procedures, which works well when tests in the individual stages are independent of each other or have the

property of p -clud. Compared to some existing methods, the proposed method has the benefits that it does not depend on any extra parameters besides the significance levels of individual stages, and it has simpler interpretation and computation, especially when the number of stages is large. It has been shown that this method works well in some applications including the one to compare two hazard rate functions of a survival data.

Generally speaking, when tests in the individual stages of a multi-stage additive testing procedure do not have the p -clud property, the proposed definition of the decision rules can not be used directly. In such cases, in order to use the proposed definition of the decision rules, some modifications of the original additive testing procedure are possible. For instance, for a two-stage procedure, let T_1 and T_2 be the test statistics in the two individual stages. When T_1 and T_2 are asymptotically, jointly, Normal distributed, then it can be checked by some simple algebraic manipulations that T_1 and $T_2 - (\rho_{12}\sigma_2/\sigma_1)T_1$ are asymptotically independent of each other, where σ_1 and σ_2 are standard deviations of T_1 and T_2 , respectively, and ρ_{12} is their correlation coefficient. Therefore, the proposed definition of the decision rules can still be used in such situations, if the test statistic used in the second-stage of the two-stage procedure is replaced by $T_2 - (\rho_{12}\sigma_2/\sigma_1)T_1$. Of course, much future research is needed, regarding estimation of σ_1 , σ_2 , and ρ_{12} , and regarding both theoretical and numerical properties of such modified procedures.

In applications, when we design a specific study, sample size determination is an important issue, about which the proposed method can be used to draw insight. To see this, let us consider a two-stage procedure once again. For a given significance level and a desired power level of the two-stage procedure for testing a specific alternative hypothesis, we can specify the corresponding significance levels and power levels for the individual tests, using similar formulas to equation (2.4). Then, sample size calculation can be done for the two individual stages. In cases for comparing two hazard rate functions, sample size calculation in the two individual stages can usually be accomplished using numerical simulations. See, for instance, Liu *et al.* (2006) for related discussion about power calculation of a test for comparing two crossing hazard rate functions.

Acknowledgments: We thank a referee for several constructive comments.

References

- Anderson, J.A., and Senthilselvan, A. (1982), “A Two-Step Regression Model for Hazard Functions,” *Applied Statistician*, **31**, 44–51.
- Bauer, P., and Köhne, K. (1994), “Evaluation of experiments with adaptive interim analysis,” *Biometrics*, **50**, 1029–1041.
- Brannath, W., Posch, M., and Bauer, P. (2002), “Recursive Combination Tests,” *Journal of the American Statistical Association*, **97**, 236–244.
- Breslow, N.E., Edler, L., and Berger, J. (1984), “A Two-Sample Censored-Data Rank Test for Acceleration,” *Biometrics*, **40**, 1049–1062.
- Cui, L., Hung, H.M.J., and Wang, S. (1999), “Modification of sample size in group sequential clinical trials,” *Biometrics*, **55**, 321–324.
- DeMets, D.L., and Ware, J.H. (1980), “Group sequential methods for clinical trials with a one-sided hypothesis,” *Biometrika*, **67** 651–660.
- DeMets, D.L., and Ware, J.H. (1982), “Asymmetric group sequential boundaries for monitoring clinical trials,” *Biometrika*, **69** 661–663.
- Fisher, R.A. (1932), *Statistical Methods for Research Workers*, London: Oliver & Boyd.
- Fleming, T.R., O’Fallon, J.R., O’Brien, P.C., and Harrington, D.P. (1980), “Modified Kolmogorov-Smirnov Test Procedures with Application to Arbitrarily Right-Censored Data,” *Biometrics*, **36**, 607–625.
- Klein, J.P., and Moeschberger, M.L. (2000), *Survival Analysis, Techniques for Censored and Truncated Data (2nd ed.)*, New York: Springer.
- Lehmacher, W., and Wassmer, G. (1999), “Adaptive sample size calculations in group sequential trials,” *Biometrics*, **55**, 1286–1290.
- Lin, X., and Wang, H. (2004), “A New Testing Approach for Comparing the Overall Homogeneity of Survival Curves,” *Biometrical Journal*, **46**, 489–496.

- Liu, K., Qiu, P., and Sheng, J. (2006), “Comparing Two Crossing Hazard Rates by Cox Proportional Hazards Modeling,” *Statistics in Medicine*, (in press).
- Liu, Q., and Chi, G. (2001), “On Sample Size and Inference for Two-Stage Adaptive Designs,” *Biometrics*, **57**, 172–177.
- Mantel, N., and Stablein, D.M. (1988), “The Crossing Hazard Function Problem,” *The Statistician*, **37**, 59–64.
- O’Brien, P.C., and Fleming, T.R. (1979), “A multiple testing procedure for clinical trials,” *Biometrics*, **35**, 549–556.
- Pampallona, S., and Tsiatis, A.A. (1994), “Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis,” *Journal of Statistical Planning and Inference*, **42**, 19–35.
- Pocock, S.J. (1977), “Group sequential methods in the design and analysis of clinical trials,” *Biometrika*, **64**, 191–199.
- Posch, M.A., and Bauer, P. (1999), “Adaptive Two-Stage Designs and the Conditional Error Function,” *Biometrical Journal*, **41**, 689–696.
- Proschan, M.A., and Hunsberger, S.A. (1995), “Designed extension of studies based on conditional power,” *Biometrics*, **51**, 1315–1324.
- Qiu, P. and Sheng, J. (2005), “A Two-Stage Procedure for Comparing Hazard Rate Functions,” (manuscript).
- Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984), “Exact Confidence Intervals Following a Group Sequential Test,” *Biometrics*, **40**, 797–803.
- Wang, S.K., and Tsiatis, A.A. (1987), “Approximately optimal one-parameter boundaries for group sequential trials,” *Biometrics* **43**, 193–199.
- Wassmer, G. (1997), “A technical note on the power determination for Fisher’s combination test,” *Biometrical Journal*, **39**, 831–838.