HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 December 27.

Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1078–1085. doi:10.1111/biom.12507.

Simultaneous Inference on Treatment Effects in Survival Studies With Factorial Designs

D. Y. Lin[1,*], **Jianjian Gong**[2], **Paul Gallo**[2], **Paul H. Bunn**[1], and **David Couper**[1]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A

[2]Novartis Pharmaceuticals Corporation, 59 Route 10, East Hanover, NJ 07936, U.S.A

SUMMARY

A clinical trial with a 2 × 2 factorial design involves randomization of subjects to treatment A or $\overline{A}$ and, within each group, further randomization to treatment B or $\overline{B}$. Under this design, one can assess the effects of treatments A and B on a clinical endpoint using all patients. One may additionally compare treatment A, treatment B, or combination therapy AB to $\overline{AB}$. With multiple comparisons, however, it may be desirable to control the overall type I error, especially for regulatory purposes. Because the subjects overlap in the comparisons, the test statistics are generally correlated. By accounting for the correlations, one can achieve higher statistical power compared to the conventional Bonferroni correction. Herein, we derive the correlation between any two (stratified or unstratified) log-rank statistics for a 2 × 2 factorial design with a survival time endpoint, such that the overall type I error for multiple treatment comparisons can be properly controlled. In addition, we allow for adjustment of prognostic factors in the treatment comparisons and conduct simultaneous inference on the effect sizes. We use simulation studies to show that the proposed methods perform well in realistic situations. We then provide an application to a recently completed randomized controlled clinical trial on alcohol dependence. Finally, we discuss extensions of our approach to other factorial designs and multiple endpoints.

Keywords

Censoring; Clinical trials; Correlated tests; Log-rank statistics; Multiple comparisons; Proportional hazards

1. Introduction

Factorial designs are commonly used in clinical trials to evaluate the effects of multiple treatments on potentially censored survival time or times to other clinical events. For example, the Physicians' Health Study adopted a 2 × 2 factorial design to investigate the effects of aspirin and beta-carotene on cardiovascular mortality and incidence of cancer among 22,000 male physicians aged 40–84 years (Stampfer et al., 1985). As a second example, the Women's Health Initiative employed a partial factorial design to study the

*D. Y. Lin, lin@bios.unc.edu.

6. Supplementary Materials

The computer code and data used in this paper are available at the *Biometrics* website on Wiley Online Libary.

effectiveness of dietary modification, hormone therapy, and calcium/vitamin D supplementation in preventing coronary heart disease, breast and colorectal cancer, and hip fracture among 68,132 postmenopausal women (Prentice and Anderson, 2007). Recently, we were involved in two clinical trials with factorial designs:

The COMBINE Study

The Combined Pharmacotherapies and Behavioral Interventions (COMBINE) study was conducted between January 2001 and January 2004 to evaluate the efficacy of medication, behavioral therapy, and their combination for treatment of alcohol dependence among 1,224 recently alcohol-abstinent volunteers (Anton et al., 2006). Patients were randomized to receive medical management with 16 weeks of naltrexone (100 mg daily) or placebo, with or without a combined behavioral intervention (CBI) under a factorial design; see Table 1. The investigators were interested in assessing the effects of each intervention as a mono-therapy, as well as the combined effect of the two interventions, on time to the first day of heavy drinking and other endpoints.

The APOLLO Trial

The Aliskiren Prevention of Later Life Outcomes (APOLLO) trial was designed to investigate the impact of aliskiren, alone or in combination with other drugs, on clinical outcomes in elderly patients with hypertension (Teo et al., 2014). Participants were randomized to aliskiren 300 mg daily or placebo and also to an additional antihypertensive drug (amlodipine 5 mg daily or HCTZ 25 mg daily) or placebo under a factorial design; see Table 2. There were two primary objectives in this study: one was to determine whether treatment with an aliskiren-based regimen reduces the risk of major cardiovascular events (i.e., death, myocardial infarction, stroke, and significant heart failure) when compared to a non-aliskiren based regimen; and the second was to determine whether intensified therapy with aliskiren plus an additional antihypertensive drug will reduce the risk of major cardiovascular events when compared to double placebo. The planned sample size was 11,000, with 2,750 patients for each of the four treatment combinations, but the trial was terminated for non-scientific reasons after enrollment of 1,759 patients (Teo et al., 2014).

In the Physicians' Health Study, the primary hypothesis pertains to the effect of aspirin on cardiovascular mortality and the secondary hypothesis pertains to the effect of beta-carotene on incidence of cancer. These are two distinct scientific questions. Likewise, the three questions considered in the Women's Health Initiative are scientifically distinct. There is no need to adjust for multiple testing in such studies (Cook and Farewell, 1996).

In the COMBINE and APOLLO studies, the same set of subjects is used to assess the effects of interventions with similar mechanisms or purposes on the same endpoint. In such cases, it may be necessary to adjust for multiple comparisons, especially when they allow different ways to make a positive claim for the benefit of an investigational treatment. Since the subjects overlap in the comparisons, the test statistics are generally dependent. By accounting for this dependence, we can achieve higher statistical power compared to the conventional Bonferroni correction. For the survival endpoint, each treatment comparison is carried out by a (stratified or unstratified) log-rank test. Because log-rank statistics are not

sums of independent random variables, it is not straightforward to determine the correlation of two log-rank statistics with overlapping subjects.

Although there is some statistical literature on 2 × 2 factorial survival experiments, no methods are available to calculate the correlations of the log-rank statistics for the types of comparisons performed in the COMBINE and APOLLO studies. The most relevant work is that of Slud (1994), who adopted the proportional hazards (PH) model (Cox, 1972) with two treatment indicators and their product as independent variables. Akritas and LaValley (1996) considered the accelerated failure time model (Kalbfleisch and Prentice, 2002, p. 44) instead of the PH model and proposed an extension of the Hodges-Lehmann estimator. Akritas and Brunner (1997) constructed nonparametric tests based on the Kaplan-Meier estimators for the four treatment combinations.

In this paper, we derive the correlation between any two (stratified or unstratified) log-rank statistics under the 2 × 2 factorial design, such that the overall type I error of multiple treatment comparisons can be properly controlled. Actually, our work goes beyond this derivation by allowing for adjustment of prognostic factors and by performing simultaneous estimation in addition to simultaneous testing. We assess the accuracy of the proposed correlation formulas in simulation studies. In addition, we apply the proposed methods to data derived from the COMBINE study. Finally, we discuss extensions of our approach to other factorial designs and multiple endpoints.

2. Methods

Consider the 2 × 2 factorial design, and let A and B denote the two treatments. Subjects are randomly assigned to A or $\overline{A}$ and B or $\overline{B}$ such that there are four possible treatment arms, as shown in the following table:

AB	$\overline{A}B$
$A\overline{B}$	$\overline{A}\,\overline{B}$

A major advantage of this design is that one can assess the effects of treatments A and B on survival time using all subjects. That is, one can assess the overall effect of treatment A by comparing the two columns in the above table or the overall effect of treatment B by comparing the two rows. One may additionally evaluate the simple effect of combination therapy AB (i.e., the comparison of the two cells on the main diagonal) or the simple effects of A and B (i.e., $A\overline{B}$ versus $\overline{A}\,\overline{B}$ and $\overline{A}B$ versus $\overline{A}\,\overline{B}$). Note that we use the term “overall effect” to refer to the effect of one intervention on survival time across the levels of the other intervention and the term “simple effect” to refer to the effect of one or two interventions on survival time compared to double placebo or standard care.

To test the null hypothesis that the overall effect of treatment A is zero, it is natural to employ the stratified log-rank statistic by stratifying subjects as to whether they receive treatment B or not (Peto, 1978). The stratified log-rank statistic compares A and $\overline{A}$ among

subjects who receive B and separately among those who receive $\overline{B}$ and then combines the evidence from the two strata. By contrast, the unstratified log-rank statistic compares all subjects who receive A with all subjects who receive $\overline{A}$ regardless of whether they receive B or $\overline{B}$. To test the null hypothesis that a simple effect is zero, we employ the unstratified log-rank statistic (since we are comparing two cells only). To quantify the treatment effect or to adjust for covariates, we appeal to the (stratified or unstratified) PH model. It is sufficient to focus on the PH model because the log-rank statistic is the score statistic under the PH model with the treatment indicator as the only independent variable. Our main task is to derive the joint distribution of the score statistics and the corresponding parameter estimators under two (possibly stratified) PH models with overlapping subjects.

Since the unstratified PH model is a special case of the stratified PH model with a single stratum, it suffices to consider the stratified PH model. Suppose that, for the first treatment comparison, there are K strata with n_k subjects in the kth stratum. In our case, $K = 1$ (for unstratified analysis) or 2 (for stratified analysis). For $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$, let T_{kj} denote the survival time for the jth subject of the kth stratum, and let $\boldsymbol{X}_{kj}$ denote the corresponding set of independent variables, including the treatment indicator (e.g., indicator of A versus $\overline{A}$ or indicator of AB versus $\overline{AB}$) and baseline covariates (e.g., age, gender, and clinical center). The stratified PH model takes the form

$$\lambda_k\left(t\middle|\boldsymbol{X}_{kj}\right)=\lambda_{k0}(t)e^{\beta^{\mathrm{T}}\boldsymbol{X}_{kj}},\quad k=1,\ldots,K;j=1,\ldots,n_k, \qquad (1)$$

where $\boldsymbol{\beta}$ is a set of regression parameters pertaining to log hazard ratios, and $\lambda_{k0}(\cdot)$ ($k = 1, \ldots, K$) are arbitrary baseline hazard functions.

For assessing the overall effect of A, it is desirable to stratify on B (Peto, 1978). Under the stratified PH model, the effects of A are assumed to be the same for subjects receiving B and for those receiving $\overline{B}$ while the difference between B and $\overline{B}$ is completely unspecified. Although the assumption of a common effect of A between the two strata may not hold (under alternatives), the stratified model is well defined and does not depend on the proportion of subjects receiving B (as opposed to $\overline{B}$). (It is possible to allow the effects of A to be different between the two strata, but then one would effectively be fitting two separate models.) Under the unstratified PH model, the baseline hazard function is a mixture of the hazard functions for $\overline{A}B$ and $\overline{AB}$, and thus its value depends on the proportion of subjects receiving B. Indeed, if the proportion of subjects receiving B is random (due to sampling or noncompliance), then the baseline hazard function and the regression effects are random quantities.

Let C_{kj} denote the censoring time for T_{kj} such that we observe $\tilde{T}_{kj} \equiv min(T_{kj}, C_{kj})$ and $\Delta_{kj} \equiv I(T_{kj} \leq C_{kj})$, where $I(\cdot)$ is the indicator function. Define

$$\boldsymbol{S}_k^{(r)}(\boldsymbol{\beta},t)=\sum_{l=1}^{n_k} I\left(\tilde{T}_{kl}\geq t\right)e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{kl}}\boldsymbol{X}_{kl}^{\otimes r},\quad r=0,1,2,$$

where $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$. We estimate $\boldsymbol{\beta}$ by maximizing the partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K}\prod_{j=1}^{n_k}\left\{\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{kj}}}{S_k^{(0)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)}\right\}^{\Delta_{kj}}.$$

The corresponding score function is

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{j=1}^{n_k}\Delta_{kj}\left\{\boldsymbol{X}_{kj} - \frac{\boldsymbol{S}_k^{(1)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)}{S_k^{(0)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)}\right\},$$

and the corresponding information matrix is

$$\mathscr{I}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{j=1}^{n_k}\Delta_{kj}\left\{\frac{\boldsymbol{S}_k^{(2)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)}{S_k^{(0)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)} - \frac{\boldsymbol{S}_k^{(1)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)^{\otimes 2}}{S_k^{(0)}\left(\boldsymbol{\beta}, \tilde{T}_{kj}\right)^2}\right\}.$$

Denote the maximum partial likelihood estimator of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$. For large samples, $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and covariance matrix $\mathscr{I}^{-1}(\hat{\boldsymbol{\beta}})$ (Andersen and Gill, 1982).

For the second treatment comparison, we consider the following stratified PH model

$$\lambda_k\left(t|\tilde{\boldsymbol{X}}_{kj}\right) = \tilde{\lambda}_{k0}(t)\, e^{\gamma^{\mathrm{T}}\tilde{\boldsymbol{X}}_{kj}}, \quad k=1,\ldots,\tilde{K}; j=1,\ldots,\tilde{n}_k, \tag{2}$$

where $\tilde{\boldsymbol{X}}_{kj}$ pertains to the treatment indicator for this comparison and baseline covariates, $\boldsymbol{\gamma}$ is a set of regression parameters, and $\tilde{\lambda}_{k0}(\cdot)\left(k=1,\ldots,\tilde{K}\right)$ are arbitrary baseline hazard functions. Estimation of model (2) proceeds in the same manner as that of model (1). Let $\hat{\gamma}$ denote the maximum partial likelihood estimator of $\boldsymbol{\gamma}$, and let $\tilde{\boldsymbol{U}}(\boldsymbol{\gamma})$ and $\tilde{\mathscr{I}}(\boldsymbol{\gamma})$ denote the score function and information matrix, respectively.

To derive the joint distribution between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, we approximate $\boldsymbol{U}(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{U}}(\boldsymbol{\gamma})$ by sums of independent terms. Specifically, $\boldsymbol{U}(\boldsymbol{\beta})$ is approximated by $\sum_{k=1}^{K}\sum_{j=1}^{n_k}\boldsymbol{w}_{kj}(\boldsymbol{\beta})$ where $\boldsymbol{w}_{kj}(\boldsymbol{\beta})$ is a random vector that involves only the data on the *j*th subject of the *k*th stratum; see equation (A.2) in the Appendix. Likewise, $\tilde{\boldsymbol{U}}(\boldsymbol{\gamma}) \approx \sum_{k=1}^{\tilde{K}}\sum_{j=1}^{\tilde{n}_k}\tilde{\boldsymbol{w}}_{kj}(\boldsymbol{\gamma})$. Replacing the unknown quantities in $\boldsymbol{w}_{kj}(\boldsymbol{\beta})$ by their sample estimators yields

$$\boldsymbol{W}_{kj}(\boldsymbol{\beta}) = \Delta_{kj}\left\{\boldsymbol{X}_{kj} - \frac{\boldsymbol{S}_k^{(1)}(\boldsymbol{\beta}, \tilde{T}_{kj})}{S_k^{(0)}(\boldsymbol{\beta}, \tilde{T}_{kj})}\right\} - \sum_{l=1}^{n_k} \frac{\Delta_{kl} I\left(\tilde{T}_{kj} \geq \tilde{T}_{kl}\right) e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{kj}}}{S_k^{(0)}\left(\boldsymbol{\beta}, \tilde{T}_{kl}\right)} \left\{\boldsymbol{X}_{kj} - \frac{\boldsymbol{S}_k^{(1)}(\boldsymbol{\beta}, \tilde{T}_{kl})}{S_k^{(0)}(\boldsymbol{\beta}, \tilde{T}_{kl})}\right\}. \quad (3)$$

Likewise, we obtain the empirical counterpart of $\tilde{w}_{kj}(\boldsymbol{\gamma})$, denoted by $\tilde{\boldsymbol{W}}_{kj}(\boldsymbol{\gamma})$. Let n_0 be the number of subjects that are used in fitting both models (1) and (2). For $i = 1, \ldots, n_0$, let $\boldsymbol{W}_i(\boldsymbol{\beta})$ be the value of $\boldsymbol{W}_{kj}(\boldsymbol{\beta})$ for the ith subject, and let $\tilde{\boldsymbol{W}}_i(\boldsymbol{\gamma})$ be the value of $\tilde{\boldsymbol{W}}_{kj}(\boldsymbol{\gamma})$ for the same subject. For large samples, the joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ is approximately multivariate normal, and the covariance matrix between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ can be estimated by $\mathscr{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{R}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\tilde{\mathscr{I}}^{-1}(\hat{\boldsymbol{\gamma}})$, where $\boldsymbol{R}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \equiv \sum_{i=1}^{n_0} \boldsymbol{W}_i(\boldsymbol{\beta})\tilde{\boldsymbol{W}}_i^{\mathrm{T}}(\boldsymbol{\gamma})$ is the estimated covariance matrix between $\boldsymbol{U}(\boldsymbol{\beta})$ and $\tilde{U}(\boldsymbol{\gamma})$.

Remark 1

If one is only interested in simple effects with a common set of covariates, then one can fit a single (unstratified) PH model with appropriate treatment indicators and obtain the covariance matrix of the estimated effects from standard statistical software.

The above results allow us to make joint inference on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Without loss of generality, assume that the first components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, denoted by β_1 and γ_1, respectively, correspond to the treatment effects. Let $\hat{\beta}_1$ and $\hat{\gamma}_1$ denote the maximum partial likelihood estimators of β_1 and γ_1, and let $\hat{\boldsymbol{\Psi}} \equiv \left\{\hat{\psi}_{l,m}; l, m = 1, 2\right\}$ denote the (estimated) covariance matrix of $(\hat{\beta}_1, \hat{\gamma}_1)$. To test jointly the null hypotheses $H_0 : \beta_1 = 0$ and $\tilde{H}_0 : \gamma_1 = 0$ we can use the quadratic form, $\left(\hat{\beta}_1, \hat{\gamma}_1\right)\hat{\boldsymbol{\Psi}}^{-1}\left(\hat{\beta}_1, \hat{\gamma}_1\right)^{\mathrm{T}}$, which is referred to the chi-squared distribution with 2 degrees of freedom. The values of β_1 and γ_1 such that $\left(\hat{\beta}_1 - \beta_1, \hat{\gamma}_1 - \gamma_1\right)\hat{\boldsymbol{\Psi}}^{-1}\left(\hat{\beta}_1 - \beta_1, \hat{\gamma}_1 - \gamma_1\right)^{\mathrm{T}} < \chi^2_{2,\alpha}$ form a joint $(1 - \alpha)$ confidence region for β_1 and γ_1, where $\chi^2_{2,\alpha}$ is the $(1 - \alpha)$ 100th percentile of the chi-squared distribution with 2 degrees of freedom.

Define $Z = \hat{\beta}_1/\hat{\psi}_{11}^{1/2}$ and $\tilde{Z} = \hat{\gamma}_1/\hat{\psi}_{22}^{1/2}$ A multiple testing procedure with an overall type I error of α is to reject H_0 if $|Z| \geq c$ and reject $\tilde{H}_0$ if $\left|\tilde{Z}\right| \geq c$, where c satisfies the equation

$$1 - Pr\left(|Z| < c, \left|\tilde{Z}\right| < c\right) = \alpha. \quad (4)$$

We evaluate this probability through multivariate normal integration, treating $\left(Z, \tilde{Z}\right)$ as bivariate zero-mean normal with unit variances and covariance $\hat{\psi}_{12}/\left(\hat{\psi}_{11}\hat{\psi}_{22}\right)^{1/2}$. The confidence intervals for β_1 and γ_1 based on c have the joint coverage probability of $(1 - \alpha)$.

Remark 2

We have implicitly assumed that the overall type I error is split equally between the two comparisons. One may spend more type I error on one hypothesis than the other by using different critical values for Z and $\tilde{Z}$ as long as the overall rejection probability satisfies equation (4).

Remark 3

We have assumed that the primary analysis involves two treatment comparisons. However, the aforementioned joint inference procedures can be easily extended to three or more comparisons since the correlation matrix is determined by the pairwise correlations.

We now consider the special case of treatment comparisons without covariate adjustment. For the first comparison, we use the stratified (weighted) log-rank statistic

$$U=\sum_{k=1}^{K}\sum_{j=1}^{n_k}\Delta_{kj}Q_k\left(\tilde{T}_{kj}\right)\left\{X_{kj}-E_k\left(\tilde{T}_{kj}\right)\right\},$$

where $E_k(t)=\sum_{l=1}^{n_k} I\left(\tilde{T}_{kl}\geq t\right)X_{kl}/\sum_{l=1}^{n_k} I\left(\tilde{T}_{kl}\geq t\right)$ and $Q_k(\cdot)$ is a possibly data-dependent weight function. The variance of U is estimated by

$$V=\sum_{k=1}^{K}\sum_{j=1}^{n_k}\Delta_{kj}Q_k^2\left(\tilde{T}_{kj}\right)\left\{E_k\left(\tilde{T}_{kj}\right)-E_k^2\left(\tilde{T}_{kj}\right)\right\}.$$

Let $\tilde{U}$ and $\tilde{V}$ be the values of U and V, respectively, for the second comparison. For the unweighted log-rank statistics (i.e., $Q_k(\cdot) = 1$), $U = U(0)$ and $\tilde{U}=\tilde{U}(0)$, such that the covariance between U and $\tilde{U}$ can be estimated by $R(0, 0)$. For non-constant weight functions, we replace Δ_{kj} and Δ_{kl} in (3) by $\Delta_{kj}Q_k\left(\tilde{T}_{kj}\right)$ and $\Delta_{kl}Q_k\left(\tilde{T}_{kl}\right)$, respectively, before evaluating $R(0, 0)$. Let $Z = U/V^{1/2}$ and $\tilde{Z}=\tilde{U}/\tilde{V}^{1/2}$. Under H_0 and $\tilde{H}_0$, $Z, \tilde{Z}$ is approximately bivariate zero-mean normal with unit variances and covariance $R(0,0)/\left(V\tilde{V}\right)^{1/2}$. This bivariate normal distribution can be used to determine the critical value c in equation (4).

It is desirable to determine the critical value c analytically, especially in the design stage. Suppose that the treatments do not affect the survival time or censoring time and that the treatment assignment ratios are 1:1 for both A versus $\overline{A}$ and B versus $\overline{B}$. We derive in the Appendix the actual values of the correlation between U and $\tilde{U}$ under various scenarios.

Specifically, for assessing the overall effect of A and the simple effect of AB, the correlation is approximately $1\sqrt{2}$. It then follows from equation (4) that $c \approx 2.1782$ for $\alpha = 0.05$. Thus, the overall type I error rate will be 0.05 if we use the nominal significance level of 0.0294 for each of the two tests. By contrast, the commonly used Bonferroni correction would entail the nominal significance level of 0.025, which is $>15\%$ smaller than 0.0294.

Remark 4

Slud (1994) considered the PH model: $\lambda\left(t|I_A, I_B\right) = \lambda_0\left(t\right) e^{\beta_1 I_A + \beta_2 I_B + \beta_3 I_A I_B}$, where I_A and I_B are indicators for treatments A and B, respectively. He obtained a closed-form expression for the covariance matrix of the maximum partial likelihood estimators of (β_1, β_2, β_3) under simplifying conditions. Note that β_1 and β_2 correspond to the simple effects of A and B, respectively, while β_3 corresponds to the interaction between A and B.

The knowledge of the critical value for each test is very useful when designing a factorial study. After the study is completed, we can obtain a more accurate value of c by empirically estimating the correlation of U and $\tilde{U}$ from the observed data, although the value of c calculated at the design stage is often accurate enough for practical purposes.

3. Simulation Studies

We conducted simulation studies to assess the performance of the proposed methods. We let the total sample size n range from 150 to 900 and randomly assigned subjects to A versus $\overline{A}$ with a 1:1 or 2:1 ratio and to B versus $\overline{B}$ with a 1:1 ratio. We generated the survival time T from the standard exponential distribution and the censoring time C from the uniform (0, 1.6) distribution such that the censoring rate is approximately 50%. We focused on the estimation of correlation for two sets of comparisons: (1) the overall effect of A and the simple effect of AB, and (2) the simple effect of A and the simple effect of AB. We investigated the accuracy of the proposed correlation estimators between the two log-rank statistics or the two maximum partial likelihood estimators. We considered both the unweighted log-rank test and the weighted log-rank test with the Kaplan-Meier estimator as the weight function.

We have shown in the Appendix that, under the treatment assignment ratios of 1:1 for A versus $\overline{A}$ and B versus $\overline{B}$, the correlation between the two log-rank statistics or the two maximum partial likelihood estimators for assessing the overall effect of A and the simple effect of AB is asymptotically $1\sqrt{2} \approx 0.707$ and the correlation for assessing the simple effect of A and the simple effect of AB is asymptotically 1/2. By extending the arguments given in the Appendix, we can show that these two correlations are $1\sqrt{2}$ and 2/3, respectively, when the treatment assignment ratio for A versus $\overline{A}$ is changed to 2:1 while that of B versus $\overline{B}$ remains at 1:1.

The results of the simulation studies are summarized in Table 3. For both the log-rank statistics and the maximum partial likelihood estimators, the empirical correlations are close to the aforementioned theoretical values. The means of the correlation estimators are slightly

below the empirical values for small n but approach the empirical values as n increases. Thus, the proposed correlation estimators are accurate enough for practical use.

4. COMBINE Study

Alcohol dependence is a leading preventable cause of morbidity and mortality and a major contributor to health care costs (Anton et al., 2006). Most patients with alcohol use disorders are never treated in primary care settings and do not receive specialty care. Although naltrexone was approved to treat alcoholism, evidence of its efficacy was based on small single-site studies using specialist models of treatment. It was of interest whether naltrexone is efficacious without specialist intervention and whether its efficacy can be improved by adding behavior therapy.

The COMBINE study was designed to assess the efficacy of naltrexone, with or without CBI, in treating alcoholism. After baseline assessment and attainment of 4 days of abstinence, 1,226 eligible alcohol-dependent individuals were randomly assigned to medical management with 16 weeks of naltrexone (100 mg daily) or placebo and were also randomly selected to receive CBI. The protocol specified percentage of days abstinent and time to first heavy drinking day (≥ 5 standard drinks per day for men, ≥ 4 for women) as two co-primary endpoints. Baseline percentage of days abstinent (within 30 days prior to the participant's last drink) and research site were prespecified covariates for both the linear and PH models. A Bonferroni-corrected significance level of 0.025 was set a priori to adjust for the two co-primary endpoints.

It is particularly important to know whether the sole act of taking naltrexone is effective given that most problem drinkers are seen in health care settings rather than in specialist treatment programs. It is also of interest whether efficacy can be improved by combining naltrexone with CBI. To assess the efficacy of naltrexone, we may combine the evidence between those who receive CBI and those who do not receive CBI or just focus on the latter group. Thus, we consider the overall effect of naltrexone, the simple effect of naltrexone, and the simple effect of naltrexone plus CBI.

Table 4 displays the results of the three comparisons. The correlation matrix for the three estimates of the log hazard ratios is

$$\begin{bmatrix} 1 & 0.673 & 0.708 \\ & 1 & 0.469 \\ & & 1 \end{bmatrix}.$$

To control the overall type I error at the prespecified level of 0.025, the critical value for the three tests is approximately 2.573. Thus, the simple effect of naltrexone is significant whereas the overall effect of naltrexone and the simple effect of naltrexone plus CBI are not. By contrast, the Bonferroni threshold for the three p-values is $0.025/3 \approx 0.0083$, which implies that none of the three tests would be significant.

The Kaplan-Meier curves shown in Figure 1 help to explain the results reported in Table 4. The use of naltrexone without CBI has the lowest likelihood of relapse. The combination therapy of naltrexone plus CBI is considerably less efficacious than the sole use of naltrexone. Thus, the overall effect of naltrexone is less significant than its simple effect. The difference between the estimates of the simple and overall effects is 0.161, with an estimated standard error of 0.10. The conclusion that the sole act of taking naltrexone is the most efficacious intervention has important clinical implications.

5. Discussion

The 2 × 2 factorial design allows one to answer multiple questions about two treatments within the same study. If the effects of one treatment are similar among patients who receive the other treatment and those who do not, then the 2 × 2 design will be efficient; otherwise, the design will reveal the complicated truth (Peto, 1978). In the COMBINE study, the effect of the sole use of naltrexone would have been estimated with more precision had we omitted CBI from the design. The factorial design, however, provided us with the opportunity to answer the question about the efficacy of combination therapy and would have provided high power to assess the overall effect of naltrexone had the effects of naltrexone been similar among patients who did or did not receive CBI.

Some literature reserves the term “factorial designs” for trials where A is tested against placebo stratifying by the level of B and B is tested against placebo stratifying by the level of A. We have defined factorial designs by the structure of the design — i.e., subjects randomly assigned to all combinations — regardless of the planned inference strategy, although we have focused on factorial designs that involve at least one stratified comparison.

Under factorial designs, which comparisons should be subject to type I error correction may be uncertain or controversial. In general, the decision depends on particular aspects of a trial, such as how scientifically distinct the comparisons are, whether they use the same endpoint, whether more than one primary hypothesis relates to the use of a single investigational treatment, and how regulatory agencies view the strategy of study sponsors. In the APOLLO trial, regulatory input was received indicating that control of the type I error across the study’s two primary hypotheses was required.

The 2×2 factorial design can be extended to the 3×3 design if each factor has three levels or to the 2 × 2 × 2 design if there are three factors of two levels each. The COMBINE study actually employed a 2 × 2 × 2 design to include a second pill, acamprosate (3 g daily), which turned out to be totally ineffective. (For the purposes of illustrating a 2 × 2 design, we did not consider the use of acamprosate in this paper.) Our general covariance formulas can be applied to any factorial design, and the arguments given in the second and third paragraphs of the Appendix can be used to obtain the specific expressions for the correlations under similar conditions.

For simplicity, we restricted our formulas to time-independent covariates. The proposed methods can be extended to time-dependent covariates in a straightforward manner. Specifically, we replace $\boldsymbol{X}_{kj}$ in model (1) and in $\boldsymbol{S}_k^{(r)}(\boldsymbol{\beta}, t)$ by $\boldsymbol{X}_{kj}(t)$, replace $\boldsymbol{X}_{kj}$ in $L(\boldsymbol{\beta})$ and

$\boldsymbol{U}(\boldsymbol{\beta})$ and in the first term of $\boldsymbol{W}_{kj}(\boldsymbol{\beta})$ given in (3) by $\boldsymbol{X}_{kj}\left(\tilde{T}_{kj}\right)$, and replace $\boldsymbol{X}_{kj}$ in the second term of $\boldsymbol{W}_{kj}(\boldsymbol{\beta})$ by $X_{kj}\left(\tilde{T}_{kl}\right)$.

In some factorial studies, there are multiple primary endpoints. As mentioned in the previous section, there were two co-primary endpoints in the COMBINE study. Because we approximated the partial likelihood score function by a sum of independent terms, the joint distribution of test statistics for any two endpoints follows from the multivariate central limit theorem. Had we accounted for the correlation between the two endpoints in the COMBINE study, we would have been able to use a less stringent nominal significance level than the Bonferroni threshold of 0.025 for each endpoint and thus to make a stronger claim about the efficacy of naltrexone.

The main contribution of this work lies in the derivation of the correlation between two treatment comparisons with overlapping subjects. The correlation is determined by the "influence function" $\boldsymbol{w}_{kj}$, which is the same ingredient for calculating the covariance matrix of the maximum partial likelihood estimators for multivariate failure time data (Wei et al., 1989). One can obtain $\boldsymbol{W}_{kj}$ in R, SAS, or STATA and perform the remaining calculations according to the formulas provided in Section 2. We have posted on our website (http://dlin.web.unc.edu/software/) a software program that estimates the correlation between any two (possibly stratified) log-rank statistics or two maximum partial likelihood estimators under (possibly stratified) PH models from the raw data. The actual values of the correlations for various scenarios derived in the Appendix are useful in the design stage and may also be accurate enough for the analysis of actual data.

There is some literature on closely related problems. Pocock et al. (1987) derived the correlation between the logrank test and a test of proportions in the context of multiple endpoints. Follmann et al. (1994) considered group sequential tests for multi-armed clinical trials. We can easily extend our work to the group sequential setting because we have approximated our statistics by sums of independent terms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the National Institutes of Health grant R01GM047845. The authors thank the Editor, an Associate Editor, and two referees for helpful comments.

Appendix: Derivation of Theoretical Results

To derive the theoretical results, we adopt the counting-process martingale formulation. Write $N_{kj}(t) = \Delta_{kj} I\left(\tilde{T}_{kj} \le t\right)$ and $Y_{kj}(t) = I\left(\tilde{T}_{kj} \ge t\right)$. Define

$$M_{kj}(t;\boldsymbol{\beta}) = N_{kj}(t) - \int_0^t Y_{kj}(u)\, e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{kj}} \lambda_{k0}(u)\, du.$$

It is easy to show that

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \int_0^{\infty} \left\{ \boldsymbol{X}_{kj} - \frac{\boldsymbol{S}_k^{(1)}(\boldsymbol{\beta}, t)}{S_k^{(0)}(\boldsymbol{\beta}, t)} \right\} dM_{kj}(t;\boldsymbol{\beta}) .$$

By the Lenglart inequality (Andersen and Gill, 1982), $\boldsymbol{U}(\boldsymbol{\beta})$ is asymptotically equivalent to

$$\sum_{k=1}^{K}\sum_{j=1}^{n_k} \boldsymbol{w}_{kj}(\boldsymbol{\beta}), \quad \text{(A.1)}$$

where

$$\boldsymbol{w}_{kj}(\boldsymbol{\beta}) = \int_0^{\infty} \{\boldsymbol{X}_{kj} - \boldsymbol{e}_k(\boldsymbol{\beta}, t)\} dM_{kj}(t;\boldsymbol{\beta}) , \quad \text{(A.2)}$$

and $\boldsymbol{e}_k(\boldsymbol{\beta}, t)$ is the limit of $\boldsymbol{S}_k^{(1)}(\boldsymbol{\beta}, t) / S_k^{(0)}(\boldsymbol{\beta}, t)$. Likewise, $\boldsymbol{U}(\boldsymbol{\gamma})$ is asymptotically equivalent to

$$\sum_{k=1}^{\tilde{K}}\sum_{j=1}^{\tilde{n}_k} \tilde{\boldsymbol{w}}_{kj}(\boldsymbol{\gamma}), \quad \text{(A.3)}$$

where $\tilde{\boldsymbol{w}}_{kj}(\boldsymbol{\gamma})$ is analogous to $\boldsymbol{w}_{kj}(\boldsymbol{\beta})$. Both (A.1) and (A.3) are sums of independent zero-mean random vectors. Thus, it follows from the multivariate central limit theorem that the joint distribution of $\boldsymbol{U}(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{U}}(\boldsymbol{\gamma})$ is asymptotically multivariate zero-mean normal with covariance matrix $\sum_{i=1}^{n_0} \boldsymbol{w}_i(\boldsymbol{\beta})\, \tilde{\boldsymbol{w}}_i^{\mathrm{T}}(\boldsymbol{\gamma})$, where $\boldsymbol{w}_i(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{w}}_i(\boldsymbol{\gamma})$ are, respectively, the values of $\boldsymbol{w}(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{w}}(\boldsymbol{\gamma})$ for the the ith subject in the overlapping set.

We now focus on the (weighted) log-rank statistics for testing the null hypotheses $H_0 : \boldsymbol{\beta} = 0$ and $\tilde{H}_0 : \boldsymbol{\gamma} = 0$. It is easy to see that $E_1(t) \approx E_2(t)$ for all t under H_0 and $\tilde{H}_0$ provided that the treatment does not differentially affect the censoring distribution between the two strata. Suppose that the same type of weight function is used for the two strata, such that $Q_1(t) \approx Q_2(t)$. Thus, we can (approximately) express U as

$$U = \sum_{i=1}^{n} \Delta_i Q\left(\tilde{T}_i\right) \left\{ X_i - E\left(\tilde{T}_i\right) \right\},$$

where $E(t) = \sum_{l=1}^{n} I\left(\tilde{T}_l \geq t\right) X_l / \sum_{l=1}^{n} I\left(\tilde{T}_l \geq t\right)$, and $n = \sum_{k=1}^{K} n_k$. Likewise,

$$\tilde{U}=\sum_{i=1}^{\tilde{n}}\Delta_i Q\left(\tilde{T}_i\right)\left\{\tilde{X}_i-\tilde{E}\left(\tilde{T}_i\right)\right\},$$

where $\tilde{E}(t)=\sum_{l=1}^{\tilde{n}} I\left(\tilde{T}_l\geq t\right)\tilde{X}_l/\sum_{l=1}^{\tilde{n}} I\left(\tilde{T}_l\geq t\right)$, and $\tilde{n}=\sum_{k=1}^{\tilde{K}}\tilde{n}_k$. Suppose that the stratification variable does not affect the survival time such that $\lambda_{10}(t)=\lambda_{20}(t)=\tilde{\lambda}_{10}(t)=\tilde{\lambda}_{20}(t)$ for all t under H_0 and $\tilde{H}_0$. (This is a reasonable approximation when designing a trial although it is theoretically possible, for example, for treatment B to have a non-zero effect even when treatment A and treatment AB have no effect.) Then simple algebraic manipulation yields

$$U=\sum_{i=1}^{n}\int_0^{\infty} Q(t)\{X_i-E(t)\}\,dM_i(t)$$

and

$$\tilde{U}=\sum_{i=1}^{\tilde{n}}\int_0^{\infty} Q(t)\left\{\tilde{X}_i-\tilde{E}(t)\right\}dM_i(t),$$

where

$$M_i(t)=N_i(t)-\int_0^t Y_i(u)\,\lambda_0(u)\,du,$$

and $\lambda_0(\cdot)$ is the common hazard function.

By the martingale central limit theorem (Andersen and Gill, 1982), U and $\tilde{U}$ are asymptotically bivariate zero-mean normal. In addition,

$$\mathrm{Var}(U)=\sum_{i=1}^{n}\int_0^{\infty} Y_i(t)\,Q^2(t)\{X_i-E(t)\}^2\lambda_0(t)\,dt,$$

$$\mathrm{Var}\left(\tilde{U}\right)=\sum_{i=1}^{\tilde{n}}\int_0^{\infty} Y_i(t)\,Q^2(t)\left\{\tilde{X}_i-\tilde{E}(t)\right\}^2\lambda_0(t)\,dt,$$

and

$$\mathrm{Cov}\left(U,\tilde{U}\right)=\sum_{i=1}^{n_0}\int_0^{\infty} Y_i(t)\,Q^2(t)\{X_i-E(t)\}\left\{\tilde{X}_i-\tilde{E}(t)\right\}\lambda_0(t)\,dt.$$

Assume that the treatment assignment ratios are 1:1 for both A versus $\overline{A}$ and B versus $\overline{B}$. Assume also that the treatments are independent of the survival time and censoring time. Then $E(t) \approx \tilde{E}(t) \approx 1/2$ for all *t*. We derive below the correlation of U and $\tilde{U}$ under various scenarios of interest.

- If U and $\tilde{U}$ pertain to the overall effect of A and the simple effect of A or AB, then $n_0 = \tilde{n}$ and $\tilde{X}_i = X_i$ for $i = 1, \ldots, n_0$. Thus, $\mathrm{Cov}\left(U, \tilde{U}\right) \approx \mathrm{Var}\left(\tilde{U}\right)$. It follows that $\mathrm{Corr}\left(U, \tilde{U}\right) \approx \left\{\mathrm{Var}\left(\tilde{U}\right)/\mathrm{Var}\left(U\right)\right\}^{1/2} \approx (\tilde{n}/n)^{1/2}$, which is $1/\sqrt{2}$ since $\tilde{n} = n/2$.
- If U and $\tilde{U}$ pertain to the overall effects of A and B, then $n_0 = \tilde{n} = n$. Because X_i and $\tilde{X}_i$ are independent, we conclude that $\mathrm{Cov}\left(U, \tilde{U}\right) \approx 0$.
- If U and $\tilde{U}$ pertain to two simple effects, then $n = \tilde{n} = 2n_0$ and $\tilde{X}_i = X_i$ for $i = 1, \ldots, n_0$. Thus, $\mathrm{Var}\left(U\right) \approx \mathrm{Var}\left(\tilde{U}\right) \approx 2\mathrm{Cov}\left(U, \tilde{U}\right)$. It follows that $\mathrm{Corr}\left(U, \tilde{U}\right) \approx 1/2$.

References

Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The Annals of Statistics. 1982; 10:1100–1200.

Anton RF, OMalley SS, Ciraulo DA, Cisler RA, Couper D, Donovan DM, et al. Combined Pharmacotherapies and Behavioral Interventions for alcohol dependence – the COMBINE study: a randomized controlled trial. Journal of the American Medical Association. 2006; 295:2003–2017. [PubMed: 16670409]

Akritas MG, LaValley MP. Nonparametric inference in factorial designs with censored data. Biometrics. 1996; 52:913–924. [PubMed: 8805761]

Akritas MG, Brunner E. Nonparametric methods for factorial designs with censored data. Journal of the American Statistical Association. 1997; 92:568–576.

Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. Journal of the Royal Statistical Society, Series A. 1996; 159:93–110.

Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B. 1972; 34:187–220.

Follmann DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. Biometrics. 1994; 50:325–336. [PubMed: 8068834]

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd. Hoboken: John Wiley & Sons; 2002.

Peto R. Clinical trial methodology. Biomedicine. 1978; 28:24–36. [PubMed: 363191]

Pocock SJ, Geller NL, Tsiatis. The analysis of multiple endpoints in clinical trials. Biometrics. 1987; 43:487–498. [PubMed: 3663814]

Prentice RL, Anderson GL. The women's Health Initiative: lessons learned. The Annual Review of Public Health. 2007; 29:131–150.

Slud EV. Analysis of factorial survival experiments. Biometrics. 1994; 50:25–38. [PubMed: 8086609]

Stampfer MJ, Buring JE, Willett W, Rosner B, Eberlein K, Hennekens CH. The 2 × 2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians. Statistics in Medicine. 1985; 4:111–116. [PubMed: 4023472]

Teo KK, Pfeffer M, Mancia G, O'Donnell M, Dagenais G, Diaz R, et al. Aliskiren alone or with other antihypertensives in the elderly with borderline and stage 1 hypertension: the APOLLO trial. European Heart Journal. 2014; 35:1743–1751. [PubMed: 24616335]

Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.

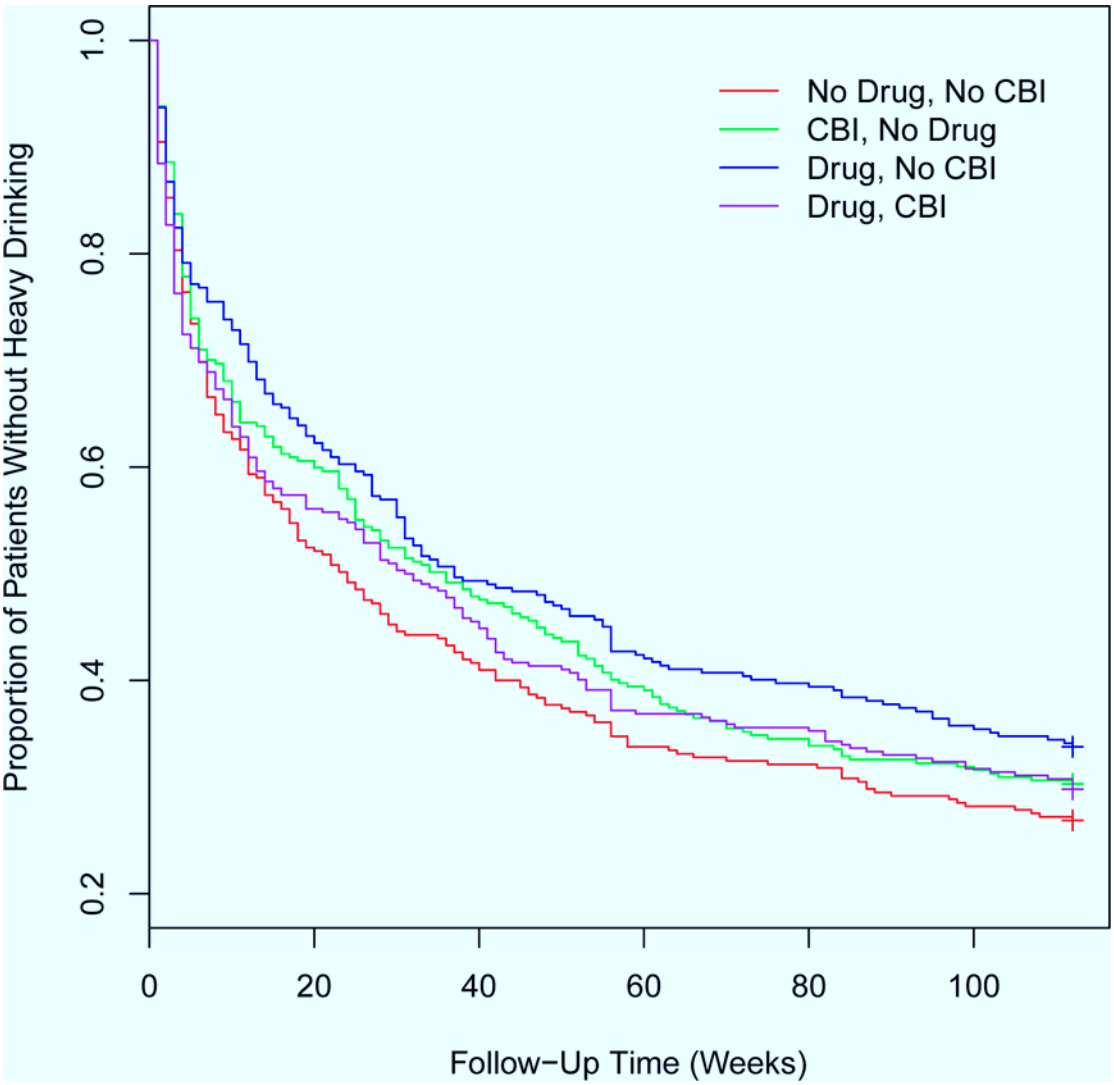

Figure 1.
Kaplan-Meier estimates of the proportion of patients without heaving drinking for the four treatment groups in the COMBINE study.

Table 1

The Factorial Design of the COMBINE Study

Naltrexone 100 mg + CBI (312 patients)	Placebo for naltrexone + CBI (307 patients)
Naltrexone 100 mg + No CBI (302 patients)	Placebo for naltrexone + No CBI (305 patients)

Table 2

The Factorial Design of the APOLLO Trial

Aliskiren 300 mg + Additional antihypertensive drug (amlodipine 5 mg or HCTZ 25 mg) (433 patients)	Placebo for aliskiren + Additional antihypertensive drug (amlodipine 5 mg or HCTZ 25 mg) (447 patients)
Aliskiren 300 mg + Placebo for additional antihypertensive drug (427 patients)	Placebo for aliskiren + Placebo for additional antihypertensive drug (452 patients)

Table 3

Simulation Results Based on 100,000 Replicates for Estimating the Correlation Between the Overall Effect of A and the Simple Effect of AB or the Correlation Between the Simple Effect of A and the Simple Effect of AB

		Unweighted Log-Rank		Weighted Log-Rank		Parameter Estimators	
A: $\overline{A}$	n	Empirical Correlation	Mean of Estimates	Empirical Correlation	Mean of Estimates	Empirical Correlation	Mean of Estimates
Overall Effect of A and Simple Effect of AB							
1:1	150	0.696	0.660	0.697	0.673	0.695	0.661
	300	0.700	0.682	0.701	0.690	0.700	0.683
	600	0.704	0.694	0.704	0.698	0.704	0.695
	900	0.707	0.698	0.707	0.701	0.707	0.699
2:1	150	0.698	0.660	0.699	0.673	0.697	0.658
	300	0.701	0.682	0.702	0.690	0.701	0.681
	600	0.703	0.694	0.704	0.698	0.703	0.694
	900	0.705	0.698	0.706	0.701	0.705	0.698
Simple Effect of A and Simple Effect of AB							
1:1	150	0.489	0.464	0.490	0.472	0.489	0.468
	300	0.495	0.482	0.495	0.488	0.495	0.484
	600	0.498	0.491	0.499	0.494	0.498	0.492
	900	0.502	0.494	0.502	0.496	0.502	0.494
2:1	150	0.656	0.622	0.658	0.635	0.667	0.623
	300	0.661	0.643	0.662	0.650	0.666	0.643
	600	0.664	0.654	0.665	0.659	0.666	0.655
	900	0.664	0.658	0.665	0.661	0.666	0.659

Table 4

Evaluation of Naltrexone With or Without CBI in the COMBINE Study

Treatment Comparison	Estimate	Std Error	*Z*-Stat	*p*-value
Overall effect of naltrexone	−0.085	0.0685	−1.237	0.216
Simple effect of naltrexone	−0.252	0.0979	−2.573	0.010
Simple effect of naltrexone + CBI	−0.091	0.0955	−0.956	0.339