Joint Testing of Overall and Simple Effects for the Two-by-Two Factorial Trial Design

Short title: Joint Testing for Two-by-Two Factorial Designs

Eric S. Leifer [1*], James F. Troendle [1], Alexis Kolecki[1], Dean A. Follmann[2]

August 24, 2019

[1]Office of Biostatistics Research, Division of Cardiovascular Sciences of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bld RLK2 Room 9206, Bethesda, MD 20892, USA

[2]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases

*Correspondence: Eric.Leifer@nih.gov, 301-523-2780

Abstract

Background/aims The two-by-two factorial design randomizes participants to receive treatment A alone, treatment B alone, both treatment A and B (AB), or neither treatment (C). When the combined effect of A and B is less than the sum of the separate A and B effects, called subadditivity, there can be low power to detect the A effect using an overall test which compares the A and AB groups to the C and B groups. Such subadditivity occurred in the Action to Control Cardiovascular Risk in Diabetes blood pressure trial (ACCORD BP) which simultaneously randomized participants to receive standard or intensive blood pressure, respectively, glycemic, control. For the primary outcome of major cardiovascular event, the overall test for efficacy of intensive blood pressure control was nonsignificant. In such an instance, simple effect tests such as A vs. C and AB vs. C may be useful since they are not affected by subadditivity, but they can have lower power since they use half the participants of the overall trial. We seek testing procedures which exploit the sample size advantage of the overall test and robustness to subadditivity of the simple tests.

Methods In the time-to-event setting, we use published asymptotic mean formulas for the stratified and simple logrank statistics to calculate the power of the overall and simple tests under various scenarios. We consider the A and B research questions to be addressing distinct scientific hypotheses and therefore allocate 0.05 significance level to each. For the A question, we propose the 2/3-1/3 procedure which allocates 2/3 of the significance level to testing the overall A effect and 1/3 to the simple AB effect. We also propose the 1/3-1/3-1/3 procedure which allocates 1/3 of the significance level to testing the overall A, simple A, and simple AB effects. These procedures are applied to the ACCORD-BP trial.

Results The 2/3-1/3 and 1/3-1/3-1/3 procedures comparably protect power better than the overall test when effects are subadditive. They have similar power to the overall test when effects are nearly additive, but only A provides benefit. For the ACCORD-BP trial primary outcome, the 2/3-1/3 and 1/3/-1/3-1/3 procedures would have detected a significant benefit of strictly controlling both blood pressure and glycemia, but not the separate effects of each.

Conclusions The 2/3-1/3 and 1/3-1/3-1/3 procedures provide reasonable power in a variety of realistic scenarios, including situations in which the overall test has reduced power. These procedures should be considered when there is uncertainty about additivity of effects and the interest is detecting any effect of the

treatments.

Keywords Factorial design, overall and simple effects, subadditivity, logrank test.

1 Introduction

The two-by-two factorial design is a popular randomized clinical trial design for simultaneously studying two experimental interventions, say, A and B. Such a design randomizes each trial participant to one of four groups: the control group C of participants who do not receive treatment A or B, group A who only receive treatment A, group B who only receive treatment B, and group AB who receive both treatments A and B. A key advantage of the factorial design over two parallel group trials of A vs. C and B vs. C is that the factorial design can asess the combined AB effect while the parallel trials cannot. Moreover, when it's reasonable to assess the overall A (respectively, B) effect against a background of 50% of the participants also receiving B (respectively, A), the factorial design uses all participants to answer the A and B research questions. Here, the overall A effect is the difference in outcomes between participants who are in the A and AB groups compared to the participants in the C and B groups; similarly for the overall B effect.

A key shortcoming to testing the overall effects is when the *simple effects* are subadditive on the scale of interest, e.g., the log hazard scale for a time-to-event outcome (Brittain and Wittes, 1989). The simple effect of A is the difference in outcomes between participants who receive A alone as compared to participants who receive C. Similarly the B simple effect and the AB simple effect, respectively, are with respect to C. The A and B effects are subadditive if the AB simple effect is less than the sum of the A and B simple effects. There are two important aspects of subadditivity. First, when the simple effects are subadditive, the simple effect of A (respectively B) does not equal the overall effect of A (respectively B). Thus, for instance, inference on the overall A effect may not apply to the simple A effect. Second, when the simple effects are subadditive, power is diminished for detecting the overall effects as compared to when the simple effects are additive. Brittain and Wittes show for normally distributed outcomes that the power loss can be substantial.

Such subadditivity occurred in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) blood pressure trial (ACCORD BP) which had 4733 high cardiovascular disease (CVD) risk, type 2 diabetes

participants. Each participant was randomized in a 50:50 fashion to receive either intensive (<120 mm Hg) or standard (<140 mm Hg) blood pressure control as well as to either intensive (< 6.0% HbA1c) or standard (7.0-7.9% HbA1c) glycemia control (Cushman, Grimm, et al. 2007). The primary outcome was the time-to-first occurrence of death due to CVD, nonfatal myocardial infarction (MI), or nonfatal stroke. Let A denote intensive blood pressure control and B denote intensive glycemia control, so C is the standard blood pressure/standard glycemia group and AB is the intensive blood pressure/intensive glycemia group. The trial design assumed a 0.80 hazard ratio (HR) for the simple A effect (which we will denote by HR_A), a 0.85 HR for the simple B effect (which we denote by HR_B), and that the simple effects were additive on the log HR scale, equivalently, multiplicative on the HR scale. Thus, if we let HR_{AB} denote the HR for the simple AB effect, it was assumed that $\log HR_A + \log HR_B = \log HR_{AB}$.

Under these assumptions, there was over 90% power to detect a significant overall A effect, i.e., conclude $HR_A^{\text{overall}} < 1$ where HR_A^{overall} denotes the overall A HR. Regardless of subadditivity, if we let $HR_{AB:B}$ denote the HR of the AB group to the B group, then by the 50:50 randomization and the assumed Cox proportional hazards model (4) (Cox, 1972):

$$\log HR_A^{\text{overall}} = \frac{\log HR_A + \log HR_{AB:B}}{2} \tag{1}$$

In words, (1) says that overall A effect which compares the A and AB groups to the C and B groups is the average of the A vs. C and AB vs. B comparisons. The overall B effect was assessed by additionally including participants from a parallel factorial trial, the ACCORD lipid trial, which randomized participants to receive intensive vs. standard glycemia control as well as intensive vs. standard lipid control. We will not consider the ACCORD lipid trial in this paper and will consider the ACCORD BP trial as if it were a stand alone trial.

For the BP trial, $HR_A^{\text{overall}} = 0.88$ was not significant (p=0.20) (Cushman, Evans, et al., 2008). However, in a *post-hoc* analysis, Margolis, O'Connor, et al. (2014) found that $HR_A = 0.74$ ($p = 0.049$), $HR_B = 0.67$ ($p = 0.011$), and $HR_{AB} = 0.71$ ($p = 0.025$). Consequently, the A and B effects were subadditive:

$$|\log HR_{AB}| = |\log(0.71)| < |\log(0.74)| + |\log(0.67)| = |\log HR_A| + |\log HR_B| \tag{2}$$

Now by (1),

$$\log(0.88) = \log HR_A^{\text{overall}} = \frac{\log HR_A + \log HR_{AB:B}}{2} = \frac{\log(0.74) + \log(0.71/0.67)}{2} \tag{3}$$

We note that $|\log HR_A^{\text{overall}}| = |\log(0.88)|$ is smaller than $|\log HR_A| = |\log(0.74)|$ because the substantial subadditivity made $\log HR_{AB:B} > 0$. Thus, power for the overall A effect was substantially diminished since it included the AB vs. B comparison.

Lin, Gong, et al. (2016) have recently explored joint testing of overall and simple effects in the Cox model setting, using the asymptotic correlation of the corresponding test statistics for computing critical values. They discuss a two-by-two factorial trial for alcoholism involving drug (A) and behavioral (B) interventions. Their interest focused on the efficacy of the drug intervention and whether it could be improved by adding behavioral therapy. Thus there was interest in the overall A effect, the simple A effect and the simple AB effect. In this paper, we also consider tests for the overall and simple effects and their statistical power. In particular, as is commonly done in factorial trials including ACCORD-BP, we allocate a two-sided 0.05 significance level to each of the A and B research questions with the idea that they are distinct questions. The remainder of this paper is organized as follows. In Section 2, we establish notation and clearly define the hypotheses of interest. In Section 3, we use Slud's (1994) and Schoenfeld's (1981) formulas for the asymptotic means of the logrank test statistics to easily derive formulas for asymptotic power and sample size. In Section 4, we propose three multiple testing procedures for the overall and/or simple effects, and explore their statistical power. As in the Brittain and Wittes paper, we are chiefly interested in subadditivity scenarios. In Section 5, we reanalyze the ACCORD-BP data with the proposed testing procedures. Finally, in Section 6, we provide further discussion and recommendations.

2 Notation and Problem

We follow the set-up from Slud (1994). Consider a 2×2 factorial trial which randomizes n participants in equal proportions to one of the four groups C, A, B, and AB. We assume there is a latent time-to-event X_i of interest for which participant i has true hazard function described by Cox's proportional hazards model:

$$\lambda(t|Z_i = (j,k)') = \lambda_0(t)\exp[j\beta_1 + k\beta_2 + jk\beta_3] \tag{4}$$

where $j, k \in \{0, 1\}$ with $j = 1$, respectively, $k = 1$, if participant i receives treatment A, respectively, B, and $\lambda_0(t)$ is the control group C's hazard function. We also assume that there is a latent censoring time V_i which is conditionally independent of X_i given Z_i.

The null hypotheses for the overall A and B effects are:

$$H_{0A}^{\text{overall}} : \beta_1 = \beta_3 = 0 \qquad \text{and} \qquad H_{0B}^{\text{overall}} : \beta_2 = \beta_3 = 0 \tag{5}$$

As is common practice (Lin, Gong, et al. 2016), we test H_{0A}^{overall} using the normalized stratified (on B) logrank statistic S_A^{str} which compares the AB group to the B group and the A group to the C group. S_A^{str} is formally defined in the Appendix. Similarly, we test H_{0B}^{overall} using the normalized stratified (on A) logrank statistic.

The null hypotheses for the simple effects of A, B, and AB are

$$H_{0A}^{\text{simple}} : \beta_1 = 0, \qquad H_{0B}^{\text{simple}} : \beta_2 = 0, \qquad H_{0,AB}^{\text{simple}} : \beta_1 + \beta_2 + \beta_3 = 0 \tag{6}$$

We test H_{0A}^{simple} using the normalized ordinary logrank statistic S_A which compares the A and C groups. Similarly we use the normalized ordinary logrank statistics S_B and S_{AB} for H_{0B}^{simple} and $H_{0,AB}^{\text{simple}}$, respectively. S_A, S_B, S_{AB} are formally defined in the Appendix.

3 Asymptotic power and sample size

We review Slud's (1994) asymptotic mean formula for the stratified logrank statistic to calculate power for tests of the overall effects. This will be used for the power studies presented in Section 4. Consider the overall A effect null hypothesis in (5) which we test using the stratified logrank statistic S_A^{str}. As discussed in the Appendix, since we assume that $\frac{n}{4}$ participants are randomized to each of the four groups, Slud's formula for the asymptotic mean of the normalized stratified logrank statistic S_A^{str} for the overall A effect is

$$\mu_A^{str} \equiv \text{asympotic mean of } S_A^{str} = (\beta_1 + 0.5\beta_3)\sqrt{0.25 \cdot n \cdot \Pr\{event\}} \tag{7}$$

Note in (7) that μ_A^{str} does not depend on the simple B effect β_2. In (7), the probability of an event is averaged over the four groups:

$$\Pr\{event\} = \frac{1}{4}\big[\Pr\{event|C\} + \Pr\{event|A\} + \Pr\{event|B\} + \Pr\{event|AB\}\big] \tag{8}$$

A heuristic way to view the asymptotic mean μ_A^{str} is by writing $\beta_1 = 0.5\beta_1 + 0.5\beta_1$. Then

$$\mu_A^{str} \equiv \text{asympotic mean of } S_A^{str} = [0.5\beta_1 + 0.5(\beta_1 + \beta_3)]\sqrt{0.25 \cdot n \cdot \Pr\{event\}} \tag{9}$$

Now β_1 is the log hazard ratio of the A group to the C group while $\beta_1 + \beta_3$ is the log hazard ratio of the AB group to the B group. With equal numbers of participants in each group, the average of β_1 and $\beta_1 + \beta_3$ is an intuitive estimate of the log hazard ratio of the AB and A groups to the B and C groups. This was demonstrated for the ACCORD BP trial results in (3).

The usual normal theory calculations show that S_A^{str} has power to reject H_{0A}^{overall} at the one-sided $\alpha/2$ significance level equal to

$$\text{power to reject } H_{0A}^{\text{overall}} = \Pr\{S_A^{str} < -z_{\alpha/2}\} = \Phi\Big(- z_{\alpha/2} - (\beta_1 + 0.5\beta_3)\sqrt{0.25 \cdot n \cdot \Pr\{event\}}\Big) \tag{10}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution and $\Phi(\cdot)$ is the standard normal distribution function. Note that in the (4) parametrization, we reject H_{0A}^{overall} in favor of a beneficial A effect for large negative values of S_A^{str}. We see in (10) that power diminishes as β_3 increases with $\beta_3 > 0$ corresponding to subadditivity of the A and B effects. The number of participants n needed to have power $1 - \beta$ to reject H_{0A}^{overall} is given by

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2}{(\beta_1 + 0.5\beta_3)^2 \Pr\{event\}} \tag{11}$$

In the Appendix, we review the power calculations for the simple effects' logrank statistics. As shown in Lin, Gong, et al. (2016), under the corresponding null hypotheses, S_A^{str} (respectively, S_B^{str}) has asymptotic correlation $1/\sqrt{2}$ with each of S_A, S_B, and S_{AB}, while the latter three simple logrank statistics have pairwise correlation 0.5 with each other.

4 Power Comparisons

4.1 Power for the Additive Effects Design

In Table 1, we use some of the design parameters from the ACCORD BP trial to examine the effect of subadditivity, and superadditivity, on the power for the overall effects. To obtain the sample size n=4160, we use an *Additive Effects Design* which assumes that the simple A and B effects are additive, as was

assumed for the ACCORD BP trial. The sample size for this design provides 90% power to detect a 0.80 hazard ratio when testing the overall A, respectively B, effect at the two-sided 0.05 significance level. Since testing simple effects may be of interest, we also include their powers. However, they are each tested at a two-sided 0.05/2=0.025 significance level since, for example, the simple A and AB effects would be relevant for establishing treatment A benefit.

We focus on the power to detect the overall A effect. The power is the highest for scenarios 1 and 2 in which the effects are additive for the originially powered HR = 0.80. Scenario 3 is also additive, but with lower power due to the lower risk reduction. Scenarios 4 and 5 correspond to low-to-moderate subadditivity, so power is lower than for scenarios 1 and 2. Scenario 6 corresponds to large subadditivity, so power is quite poor. Power is better for scenario 7 despite $HR_A = 0.90$ since the effects are superadditive. The last scenario corresponds to the observed HRs in the ACCORD-BP trial. Since the effects have large subadditivity, there is low *post hoc* power, 41%, for the overall A effect (intensive BP control) which was the trial's primary analysis. It is useful to note in Table 1 that across the scenarios, some combination of the overall A, simple A, and simple AB effects have reasonable power. This is explored in the next section.

4.2 Multiple testing procedures for treatment A benefit

If we are interested in detecting an A benefit, whether in combination with B or not, there are various testing procedures which could be considered, four of which are presently discussed. Since we consider the A and B research questions to be distinct, type I familywise error (FWE) control will be with respect to the hypotheses involving A (respectively, B). Thus, for A, we require FWE control at the 0.05 level for any combination of the hypotheses H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$.

Lin, Gong, et al. (2016) discussed the possibility of allocating the FWE among the overall and simple hypotheses in various ways. In Table 2, we examine four different procedures at the modestly (11%) increased sample size n=4600. All of the calculations use the `fac2x2design` function from the `factorial2x2` R package (Leifer and Troendle, 2019). The overall A test at the 0.05 significance level is the same as in Table 1, that is, all of the FWE is allocated to the overall A test. The *2/3-1/3* procedure jointly tests the overall A and the simple AB effects. The overall A effect is tested at the two-sided $(2/3) \cdot 0.05$ level corresponding to a critical

value $=-2.13$. We then use the $1/\sqrt{2}$ asymptotic correlation between the overall A and simple AB logrank statistics to obtain a critical value $= -2.24$ for the simple AB test which controls the FWE at the 0.05 level. Details of the calculation of -2.24 are given in the Appendix. We note that -2.24 corresponds to a two-sided 0.0251 significance level which is larger than the Bonferroni-corrected $0.05/3 = 0.0167$ significance level which would have been used had we not exploited the correlation. The *1/3-1/3-1/3* procedure jointly tests the overall A, simple A, and simple AB effects using a common critical value. We use the $1/\sqrt{2}$ asymptotic correlation between the overall A and simple AB logrank statistics as well as the $1/2$ asymptotic correlation between the simple A and simple AB statistics to calculate the common critical value $=-2.32$ for each of the three tests which controls the FWE at the 0.05 level. We note that -2.32 corresponds to a two-sided significance level of 0.0203 which is larger than the Bonferroni-corrected $0.05/3 = 0.0167$ significance level. Finally, the *1/2-1/2* procedure jointly tests the simple A and simple AB effects. We use the $1/2$ asymptotic correlation between the simple A and simple AB statistics to calculate a common critical value $= -2.22$ which controls the FWE at the 0.05 level. We note that -2.22 corresponds to a two-sided significance level of 0.0264 which is slightly larger than the Bonferroni-corrected $0.05/2 = 0.025$ significance level.

For the Table 2 scenarios, the 2/3-1/3 and 1/3-1/3-1/3 procedures are the most robust with respect to power for detecting an A effect. Indeed, for scenario 6 and the ACCORD BP scenario where there is substantial subadditivity, the overall A test has 41-45% power. The 2/3-1/3 and 1/3-1/3-1/3 procedures' powers are substantially better for those scenarios because the simple A and/or AB effects are also tested, and power for those simple effects are unaffected by subadditivity. In contrast, for scenario 3 which has additive effects, the 1/2-1/2 design has only 52% power since it only tests the simple A and AB effects. The 2/3-1/3 and 1/3-1/3-1/3 procedures' powers are noticeably better since they also test the overall A effect and there is more power to detect the overall effect than the simple effects in this additive effects scenario.

5 Re-analysis of ACCORD-BP intensive blood pressure control effects

Table 3 summarizes a re-analysis of the ACCORD-BP trial's intensive blood pressure control effects. To be consistent with the primary results paper (Cushman, Evans, et al. 2008), the Cox models on which the results are based adjust for the presence or absence of a previous cardiovascular event, as well as the clinical center network to which the participant was enrolled. Also, as was done for the primary results, participants were followed for a maximum of 7 years. This is why the hazard ratio estimates and p-values are slightly different from the Margolis, O'Connor, et al. (2014) results. Those results censored follow-up no later than the time at which the intensive glycemic control participants (in the combined ACCORD BP and ACCORD lipid trials) were informed of the data safety monitoring board's recommendation to discontinue intensive glycemic control (Gerstein, Miller, et al. 2011).

Since the ACCORD BP Cox models adjust for baseline covariates, the nominal significance levels for rejection for the 2/3-1/3, 1/3-1/3-1/3, and 1/2-1/2 procedures are slightly different from those reported in the Table 2 legend. This is because the Table 2 significance levels depend on the asymptotic correlations between the *unadjusted* logrank statistics for the overall A, simple A, and simple AB tests. To take into account the adjustment covariates, we use the methodology of Lin, Gong, et al. (2016) to calculate the correlations of the corresponding log hazard ratios. We use the `fac2x2analyze` function from the `factorial2x2` R package (Leifer and Troendle, 2019) to calculate the correlation matrix to be:

$$\begin{array}{r} \text{overall } BP \text{ effect} \\ \text{simple } BP \text{ effect} \\ \text{simple } BP+ \text{ glycemia effect} \end{array} \begin{pmatrix} 1 & 0.733 & 0.728 \\ & 1 & 0.426 \\ & & 1 \end{pmatrix} \tag{12}$$

We see that for the ACCORD BP data, the correlation between the covariate-adjusted overall and simple effect estimates, 0.733 and 0.728, respectively, are slightly larger than the asymptotic correlation $1/\sqrt{2} = 0.707$ between the unadjusted overall and simple estimates. On the other hand, the correlation between the covariate-adjusted simple effects, 0.426, is smaller than the asymptotic correlation $1/2 = 0.5$ between the unadjusted simple estimates. Using the above correlation matrix, `fac2x2analyze` calculates that the

2/3-1/3 procedure tests the overall A effect at the 0.033 level and the simple AB effect at the 0.02605 level; the 1/3-1/3-1/3 procedure tests each of its effects at the 0.0210 level; and the 1/2-1/2 procedure tests each of its effects at the 0.0264 level. As seen in Table 3, the only significant effect is the simple BP + glycemia effect, which is significant for all three multiple comparison procedures.

6 Discussion

For a two-by-two factorial trial with a time-to-event endpoint, we used Slud's (1994) asymptotic results to study the effect of subadditivity on the power to detect the overall A and B effects. We assumed a true Cox proportional hazards model with independent censoring. We first considered the Additive Effects Design which assumes the simple A and B effects are additive. It has 90% power to test the overall A, respectively, B, effect at the two-sided 0.05 significance level. Such a design considers the A and B research questions to be distinct, which is the context of this paper. As long as there is mild subadditivity, there is good power for the overall A effect regardless of the simple B effect. However, as demonstrated by Wittes and Brittain (1989) for normally distibuted outcomes, moderate subadditivity causes a dramatic dimunition in the power to detect the overall effects. Nevertheless, when there is moderate subadditivity, there can be reasonable power to detect the simple AB effect if it has a greater HR compared to the control group than either simple A or simple B.

The last observation led us to consider for the A research question whether joint testing of various meaningful combinations of the overall effect and simple effects could provide reasonable power across the scenarios we examined. To do so, we proposed three procedures which require an 11% sample size increase for the parameter settings we used for the Additive Effects Design. The 2/3-1/3 procedure allocates 2/3 of the 0.05 significance level to the overall A test and 1/3 of the signficance level to the simple AB test. The 1/3-1/3-1/3 procedure allocates 1/3 of the significance level to each of the overall A, simple A, and simple AB tests. The 1/2-1/2 procedure allocates 1/2 of the significance level to each of the simple A and simple AB tests. Across the scenarios we considered, the 2/3-1/3 and 1/3-1/3-1/3 procedures were more robust in the mini-max power sense for detecting some A effect than the 1/2-1/2 procedure or the overall A test. This is because the 2/3-1/3 and 1/3-1/3-1/3 procedures each jointly test the overall effect and simple effects while

the 1/2-1/2 procedure only tests the simple effects. Only testing the simple effects does not take advantage of using all of the participants for answering the research question.

We have presented joint testing procedures for the A research question, controlling the FWE at the 0.05 level for the A-centric hypotheses of interest. Analogous joint testing procedures would be performed for the B research question. controlling the FWE at the 0.05 level for the B-centric hypotheses. In particular, for proper inference, the simple A, B, and AB effects should be examined. For example, consider the scenario $HR_A = 1.0$, $HR_B = 0.8$, and $HR_{AB} = 0.8$ so the entire AB effect is through B. Under Table 2's parameter settings, the 2/3-1/3 and 1/3-1/3-1/3 procedures for the A research question have respective powers of 61.3% and 58.2% because of the simple AB effect. Examination of the simple A and B effects would very likely suggest that B has a larger benefit since there is a greater than 99% probability that the estimated log hazard ratio for B will be more negative than the estimated log hazard ratio for A . Moreover, by switching the A and B labels in Scenario 2 in Table 2, the B-centric 2/3-1/3 and 1/3-1/3-1/3 procedures have respective powers of 93.1% and 91.1%. Thus, there would be high power to declare the importance of B.

Of course the research question should motivate the testing procedure at least as much as power considerations. In the alcohol treatment trial described by Lin, Gong, et al. (2016) which was discussed in the Introduction, the 1/3-1/3-1/3 would have been most appropriate for their research question. They presented the results of their 1/3-1/3-1/3 analysis in their paper.

A separate issue is whether the A and B research questions are distinct or not with respect to controlling type I error. Korn and Freidlin (2016) have proposed a simple three-step procedure with sequentially ordered simple effects' hypotheses which controls the familywise type I error across all of the A and B comparisons. Since Korn and Freidlin only consider simple effects' tests, subadditivity is not a concern. For the Table 1 parameter settings, their procedure requires about 65% more participants. This is because each of the simple effects A, B, and AB have 80% power when tested against C at the 0.05/3 level at the first step of their procedure. In subsequent steps, other comparisons may be tested, such as simple A vs. simple B, as well as simple AB vs. the best of C, A, or B. The Korn-Freidlin procedure should certainly be considered if it is felt that familywise type I error should be controlled at the 0.05 level across the combined families of A and B hypotheses, substantial subadditivity may exist, and the larger required sample size is feasible.

In conclusion, the two-by-two factorial design can be a very useful design for simultaneously answering several research questions. However, care must be exercised in choosing the testing procedure to be used. Several things need to be considered including a clear articulation of the research questions of interest, whether the A and B research questions may be considered distinct, and the extent to which subadditivity is a concern. When the research questions are distinct, the 2/3-1/3 and 1/3-1/3-1/3 procedures are useful as they exploit the sample size advantage of the overall test while incorporating simple effect testing which is not affected by subadditivity.

Declaration of conflicting interests

The views expressed in this paper are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; National Institutes of Health; or the United States Department of Health and Human Services. The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

References

1. Brittain E and Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Statistics in Medicine*. 1989; 8: 161-171.

2. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*. 1972; 34: 187-219.

3. Cushman WC, Grimm RH Jr, Cutler JA, et al; ACCORD Study Group. Rationale and design for the blood pressure intervention of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial. *Am J Cardiol*. 2007; 99(12A):44i-55i.

4. Cushman WC, Evans GW, Byington RP, et al.; ACCORD Study Group. Effects of intensive blood-

pressure control in type 2 diabetes mellitus. *N Engl J Med*. 2008; 358: 2545-2559.

5. Gerstein HC, Miller ME, Genuth S, et al.; ACCORD Study Group. Long-term effects of intensive glucose lowering on cardiovascular outcomes. *N Engl J Med*. 2011;364:818-828.

6. Korn EL and Freidlin B. Non-factorial analyses of two-by-two factorial designs. *Clinical Trials*. 2016; 13: 651-659.

7. Leifer ES and Troendle JF. factorial2x2: Design and Analysis of a 2x2 factorial trial with a time to event endpoint. R package version 0.1. 2019. URL https://github.com/EricSLeifer/factorial2x2.

8. Lin D-Y, Gong J, Gallo P, et al. Simultaneous inference on treatment effects in survival studies with factorial designs. *Biometrics*. 2016; 72: 1078-1085.

9. Margolis, O'Connor, et al. Outcomes of combined cardiovascular risk factor management strategies in type 2 diabetes: the ACCORD randomized trial. *Diabetes Care*. 2014; 37: 1721-1728.

10. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 1981; 68: 316-319.

11. Slud EV. Analysis of factorial survival experiments. *Biometrics*. 1994; 50: 25-38.

7 Appendix

7.1 Stratified and simple logrank statistics

We define the stratified and simple logrank statistics used in this paper as in Slud (1994). Using the notation from Section 2, let $T_i = \min(X_i, V_i)$ be participant's i's observed follow-up time and $\Delta_i = I(X_i \leq V_i)$ be the indicator that participant i had an observed event. For $j, k = 0, 1$, let $Y_{jk}(t) = \sum_{i=1}^{n} I[T_i \geq t, Z_i = (j,k)']$ be the at-risk set for stratum (j,k) at time t. Then the normalized stratified logrank statistic S_A^{str} is defined as

$$S_A^{\text{str}} = \frac{\sum_{i=1}^{n} \Delta_i \sum_{k=0}^{1} \left\{ I[Z_i = (1,k)] - \frac{Y_{1k}(X_i)}{Y_{1k}(X_i) + Y_{0k}(X_i)} I[Z_i^{(2)} = k] \right\}}{\sqrt{\sum_{i=1}^{n} \Delta_i \sum_{k=0}^{1} \frac{Y_{1k}(X_i) Y_{0k}(X_i)}{Y_{1k}(X_i) + Y_{0k}(X_i)} I[Z_i^{(2)} = k]}} \tag{13}$$

where $Z_i^{(2)}$ is the second component of Z_i. S_B^{str} is similarly defined.

To obtain the asymptotic mean of S_A^{str} in (7), we use Table 1 in Slud (1994). In particular, $c_1 = \sqrt{n}\beta_1$, $c_3 = \sqrt{n}\beta_3$, $D_{11} = 0.5 \cdot 0.5 \cdot \Pr\{event\} = 0.25 \cdot \Pr\{event\}$, $D_{13} = 0.5 \cdot 0.5 \cdot 0.5 \cdot \Pr\{event\} = 0.125 \cdot \Pr\{event\}$.

The normalized simple logrank statistic S_A is defined as

$$S_A = \frac{\sum_{i=1}^n \Delta_i I[Z_i^{(2)} = 0]\left\{ I[Z_i = (1,0)] - \dfrac{Y_{10}(X_i)}{Y_{10}(X_i) + Y_{01}(X_i)} \right\}}{\sqrt{\sum_{i=1}^n \Delta_i I[Z_i^{(2)} = 0] \dfrac{Y_{10}(X_i) Y_{00}(X_i)}{Y_{10}(X_i) + Y_{00}(X_i)}}} \tag{14}$$

S_B and S_B are similarly defined. Since $\dfrac{n}{4}$ participants are randomized to each of the A and C groups, the results from Schoenfeld (1981) show that S_A has asymptotic mean

$$\mu_A = \beta_1 \sqrt{0.25 \cdot \frac{n}{2} \cdot \Pr\{event | A \cup C\}} \tag{15}$$

The usual normal theory calculations show that S_A has power to reject H_{0A}^{simple} at the one-sided $\alpha/2$ significance level equal to

$$\text{power to reject } H_{0A}^{\text{simple}} = \Pr\{S_A < -z_{\alpha/2}\} = \Phi\left(-z_{\alpha/2} - \beta_1 \sqrt{0.25 \cdot \frac{n}{2} \cdot \Pr\{event | A \cup C\}} \right) \tag{16}$$

7.2 Calculating the critical values and power for the 2/3-1/3 procedure

First we calculate the critical values for the 2/3-1/3 procedure whose FWE is controlled at the two-sided 0.05 significance level. We test the overall A effect using the stratified logrank statistic S_A^{str} at the two-sided $(2/3) \cdot 0.05$ significance level which corresponds to a -2.13 critical value. We test the simple AB effect using the simple logrank statistic S_{AB}. We reject the null hypotheses H_{0A}^{overall} and $H_{0,AB}^{\text{simple}}$ for large negative values of S_A^{str} and S_{AB}, respectively. Assuming both null hypotheses H_{0A}^{overall} and $H_{0,AB}^{\text{simple}}$ are true, $(S_A^{\text{str}}, S_{AB})$ are jointly asymptotically normal each with mean vector $(0,0)$, variance 1, and correlation $1/\sqrt{2}$. The critical value c for the simple AB effect is the solution to

$$0.025 = \Pr\{[S_A^{\text{str}} < -2.13] \cup [S_{AB} < c]\} = 1 - \Pr\{[S_A^{\text{str}} > -2.13] \cap [S_{AB} > c]\} \tag{17}$$

We use the `crit2x2` function from the `factorial2x2` R package to calculate $c = -2.24$ (Leifer and Troendle, 2019).

As an example of a power calculation for the 2/3-1/3 procedure, consider the 81.2% power in Table 2, scenario 4. To compute the asymptotic mean μ_A^{str} of S_A^{str} in (7), we have $\beta_1 = \log(0.80)$, $\beta_3 = \log\left(\frac{0.95}{0.80 \cdot 1.10}\right)$, $n = 4600$, and $\Pr\{event\} = 0.236$, so $\mu_A^{str} = -3.046$. To compute the asymptotic mean of μ_{AB} of S_{AB} analogously to (15), we have $\beta_1 + \beta_2 + \beta_3 = \log(0.95)$ and $\Pr\{event|AB \cup C\} = 0.239$ so $\mu_{AB} = -0.601$. The asymptotic variance of S_A^{str} and S_{AB} is 1 and asymptotic correlation is approximately $1/\sqrt{2}$. Then under those mean, variance, and correlation parameters, we use the `fac2x2design` function from the `factorial2x2` package to compute

$$\Pr\{[S_A^{\text{str}} < -2.13] \cup [S_{AB} < -2.24]\} = 0.821. \tag{18}$$

It is important to note that under the local alternatives in scenario 4, as opposed to the intersection of the null hypotheses $H_{0A}^{\text{overall}} \cap H_{0,AB}^{\text{simple}}$, the asymptotic correlation of S_A^{str} and S_{AB} is a few absolute percentage points away from $1/\sqrt{2}$. However, this affects the power by less than one absolute percentage point, so the analytical power calculation using `fac2x2design` is safe and easier to use than a simulated power calculation.

The critical values and power of the 1/3-1/3-1/3 and 1/2-1/2 procedures are obtained in a similar manner.

Table 1: Power for Additive Effects Design using (10) and (16): $\boldsymbol{n}$**=4160***.

Scenario	True HRs vs. C				Power (%) to declare a statistically signficant benefit for:				
				Level	0.05	0.05	0.025	0.025	0.025
	HR_A	HR_B	HR_{AB}		overall $A^\natural$	overall $B^\natural$	simple $A^\sharp$	simple $B^\sharp$	simple $AB^\sharp$
1	0.80	0.80	0.64		90.0	90.0	56.4	56.4	99.1
2	0.80	1.00	0.80		90.0	2.5	56.4	1.2	56.4
3	0.85	1.00	0.85		70.7	2.5	31.9	1.2	31.9
4	0.80	1.10	0.95		82.6	0.0	56.4	0.0	4.8
5	0.80	0.80	0.72		67.5	67.5	56.4	56.4	88.9
6	0.80	0.80	0.80		38.1	38.1	56.4	56.4	56.4
7	0.90	0.90	0.72		69.7	69.7	14.0	14.0	89.1
ACCORD-BP	0.74	0.67	0.71		41.1	88.4	82.9	97.7	91.3

*True hazard function is (4) with a 4.45% annual event rate in the control group so $\lambda_0(t) = 0.0455$. Independent Uniform(4.0, 8.4) years censoring, so expected number of events = 816-983, depending on the scenario. Assuming the hazard ratios $HR_A = HR_B = 0.80$ are additive on the log scale, i.e., $HR_{AB} = 0.64$, there is 90% power for the overall A (respectively, B) effect at the two-sided 0.05 significance level.

$^\natural$ Tests H_{0A}^{overall} and H_{0B}^{overall} each at the two-sided 0.05 significance level.

$^\sharp$ Tests H_{0A}^{simple}, H_{0B}^{simple}, $H_{0,AB}^{\text{simple}}$ each at the two-sided 0.025 significance level.

Table 2: Power for joint testing procedures for A: $\boldsymbol{n}$**=4600***

Scenario	True HRs vs. C			Power (%) to declare a statistically significant benefit for			
	HR_A	HR_B	HR_{AB}	Overall $A^\natural$	2/3-1/3†	1/3-1/3-1/3$^\flat$	1/2-1/2$^\sharp$
1	0.80	0.80	0.64	92.6	99.6	99.5	99.6
2	0.80	1.00	0.80	94.7	93.1	91.1	77.8
3	0.85	1.00	0.85	75.0	71.2	67.3	51.6
4	0.80	1.10	0.95	86.2	82.1	80.8	62.4
5	0.80	0.80	0.72	71.8	92.9	93.0	94.1
6	0.80	0.80	0.80	41.4	64.6	74.8	77.9
7	0.90	0.90	0.72	74.0	93.0	91.7	92.4
ACCORD BP	0.74	0.67	0.71	44.7	94.0	96.7	97.4

*True hazard function is (4) with a 4.45% annual event rate in the control group so $\lambda_0(t) = 0.0455$. Independent Uniform(4.0, 8.4) years censoring so expected number of events = 902-1087 depending on the scenario.

$^\natural$ Tests H_{0A}^{overall} at the two-sided 0.05 significance level

† Tests H_{0A}^{overall} at the two-sided $(2/3)\cdot 0.05$ level and $H_{0,AB}^{\text{simple}}$ at the two-sided 0.0251 level giving 0.05 FWE

$^\flat$ Tests H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$ each at the two-sided 0.0203 level giving 0.05 FWE

$^\sharp$ Tests H_{0A}^{simple} and $H_{0,AB}^{\text{simple}}$ each at the two-sided 0.0264 level giving 0.05 FWE

Table 3: Re-analysis of intensive blood pressure control with or without intensive glycemic control in the ACCORD-BP trial

Treatment comparison	HR (95% CI)*	P-value*
Overall effect of strict BP control	0.88 (0.73-1.06)	0.194
Simple effect of strict BP control	0.76 (0.59-0.98)	0.035
Simple effect of strict BP and glycemic control	0.73 (0.56-0.94)	0.016$^\natural$

* Adjusted for the presence of a previous cardiovascular event and the clinical center network to which the participant was enrolled

$^\natural$ Significant with respect to the 2/3-1/3, 1/3-1/3-1/3, and 1/2-1/2 procedures.