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]

April 1, 2020

[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 a subadditve interaction, there can be low power to detect the A effect using an overall test, i.e., factorial analysis, which compares the A and AB groups to the C and B groups. Such an interaction may have been present in the Action to Control Cardiovascular Risk in Diabetes blood pressure trial (ACCORD BP) which simultaneously randomized participants to receive intensive or standard 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 of A vs. C and B vs. C may be useful since they are not affected by a subadditive interaction, but they can have lower power since they use half the participants of the overall trial. We investigate multiple testing procedures which exploit the overall tests' sample size advantage and the simple tests' robustness to a potential interaction.

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 unrelated hypotheses and therefore allocate 0.05 significance level to each. For each question, we investigate three multiple testing procedures which allocate the type 1 error in different proportions for the overall and simple effects as well as the AB effect. The Equal Allocation 3 procedure allocates equal amounts of the type 1 error to testing each of the three effects, the Proportional Allocation 2 procedure allocates 2/3 of the type 1 error to testing the overall A (respectively, B) effect and the remaining type 1 error for the AB effect, and the Equal Allocation 2 procedure allocates equal amounts to testing the simple A (respectively, B) and AB effects. These procedures are applied to the ACCORD-BP trial.

Results Across various realistic clinical trial scenarios with interactions of different magnitudes, the Equal Allocation 3 procedure had robust power for detecting a true beneficial effect. For the ACCORD-BP trial all three procedures would have detected a significant benefit of strict glycemia control.

Conclusions The Equal Allocation 3 procedure provides robust power across many scenarios, including when the overall test has reduced power due to an interaction.

Keywords Interaction, subadditive, logrank test, Dunnett correction.

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. In this paper, we will assume that one-fourth of the participants are randomized to each of the four groups. The advantages of the factorial design over two parallel group trials of A vs. C and B vs. C are that the factorial design can assess assess the A and B effects within the same trial as well as the combined AB effect.

There are two types of effects which are of interest in a factorial design: *simple* and *overall*. 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, respectively, AB, simple effects are the comparisons of B, respectively, AB, versus C. The overall A effect, sometimes called the main or factorial 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, the overall B effect compares the B and AB participants to the C and A participants. Tests of the overall effects correspond to the standard factorial analysis for many factorial trials.

The overall A effect is of interest when there is no A-by-B interaction since then it corresponds to the A vs. C difference, as does the simple A effect. In the absence of an interaction, the overall A effect can be more precisely estimated, and tested with greater power, than the simple A effect since the overall A effect is estimated using all of the trial's participants while the simple A effect is estimated using half of the trial's participants.

To see why the overall A effect corresponds to the A vs. C difference when there is no interaction, suppose A and B corresponded to different blood pressure medications which reduced blood pressure by 5 and 7 mm Hg, respectively. Further, suppose C was a placebo group and had no blood pressure reduction. When there is no interaction, the combination of the A and B medications, i.e., AB, would reduce blood pressure by 12 mm Hg. In other words, the A and B effects are additive.. In this case, the overall A effect corresponds to the A vs. C difference of a 5 mm Hg reduction since 5 mm Hg is the difference between both the A and C groups ($5 - 0$ mm Hg) and the AB and B groups ($12 - 7$ mm Hg). Similar comments apply to the overall and simple B effects.

However, it can be the case that A and B have a subadditive interaction which, in our blood pressure example, would mean that the combination of the A and B medications reduce blood pressure by less than 12 mm Hg. A subadditive interaction reduces the power for detecting an overall effect, as compared to when there is no interaction. It also complicates the interpretation of an overall effect since it no longer equals the simple effect (Brittain and Wittes, 1989).

It is also possible for A and B to have a superadditive interaction which in our blood pressure example would mean their combination reduces blood pressure by more than 12 mm Hg. In this case, there is greater power for detecting an overall effect than if there were no interaction. However, since diminished power is of a greater concern, our focus will be on subadditive interactions, which we will simply refer to as interactions.

There is rarely a pre-trial way to know whether an interaction exists. Moreover, there are few factorial trials which are adequately powered to detect an interaction since that would require approximately quadrupling the sample size required to detect an overall effect assuming no interaction (Peterson and George, 1993). Nevertheless, the no-interaction assumption may be reasonable in the prevention setting when toxicities are negligible (Freidlin and Korn, 2017). We now discuss in detail an important cardiovascular prevention factorial trial which assumed no interaction existed for its power calculations for the overall effect, but for which there was some post-trial evidence of an interaction.

The Action to Control Cardiovascular Risk in Diabetes blood pressure trial (ACCORD BP) trial enrolled 4733 high cardiovascular disease (CVD) risk, type 2 diabetes participants. Each trial 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. Assuming Cox's proportional hazards model (Cox, 1972), the trial had over 90% power to detect an overall blood pressure effect at the two-sided 0.05 significance level. The power calculation assumed a 0.80 hazard ratio (HR) for the simple blood pressure effect (which we will denote by HR_A), a 0.85 HR for the simple glycemia effect (which we denote by HR_B), and that there was no interaction. Thus, if we let HR_{AB} denote the HR for the simple blood pressure + glycemia effect, it was assumed that $\log HR_A + \log HR_B = \log HR_{AB}$.

To explore the ACCORD BP trial results, let $\widehat{HR}_A^{\text{overall}}$ denote the estimated overall blood pressure effect HR and let $\widehat{HR}_{AB:B}$ denote the estimated HR of the blood pressure + glyecmia group to the glycemia group. By the 50:50 randomization and the assumed Cox proportional hazards model in equation (4):

$$\log \widehat{HR}_A^{\text{overall}} = \frac{\log \widehat{HR}_A + \log \widehat{HR}_{AB:B}}{2} \tag{1}$$

In words, (1) says that estimated overall blood pressure effect which compares the A and AB groups to the C and B groups is the average of the estimated A vs. C and AB vs. B comparisons. The overall glycemia 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 ACCORD BP trial, the primary analysis was based on the estimated overall blood pressure effect hazard ratio which was $\widehat{HR}_A^{\text{overall}} = 0.88$ and not significantly different from 1 (p=0.20) (Cushman, Evans, et al., 2008). However, in a *post-hoc* analysis, Margolis, O'Connor, et al. (2014) found that the estimated simple BP effect hazard ratio $\widehat{HR}_A = 0.74$ was significant at the 0.05 significance level with $p = 0.049$. Thus, basing the primary analysis on the overall BP effect instead of the simple BP effect may have resulted in missing an opportunity to declare the efficacy of strict blood pressure control. Moreover, Cushman, Evans, et al. found that this missed opportunity may have occurred because of an interaction. Indeed, they estimated the simple glycemia effect to be $\widehat{HR}_B = 0.67$ ($p = 0.011$), and the simple blood pressure + glycemia effect to be $\widehat{HR}_{AB} = 0.71$ ($p = 0.025$). Consequently, although the test for an interaction was not significant ($p = 0.08$), the estimated simple AB HR was less in absolute terms than the sum of the simple A and B HRs:

$$|\log \widehat{HR}_{AB}| = |\log(0.71)| = 0.342 < 0.702 = |\log(0.74)| + |\log(0.67)| = |\log \widehat{HR}_A| + |\log \widehat{HR}_B| \tag{2}$$

Now by (1),

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

We note that the estimated overall blood pressure effect $|\log \widehat{HR}_A^{\text{overall}}| = |\log(0.88)| = 0.128$ is smaller than the simple blood pressure effect $|\log \widehat{HR}_A| = |\log(0.74)| = 0.301$. This is because the estimated simple blood pressure + glycemia vs. simple glycemia log hazard ratio $\log \widehat{HR}_{AB:B} = \log(0.71/0.67) > 0$. Thus, power for the overall blood pressure effect may have been diminished since it included the blood pressure + glycemia (AB) vs. glycemia (B) comparison.

The ACCORD BP experience suggests that we construct multiple testing procedures which ideally have very good power for detecting the overall A effect, which use all the participants, when there is no interaction, and reasonable power for the simple A and AB effects, each of which use half of the participants, when there is an interaction (similar comments apply to the overall B and simple B effects). This is an idea explored by Lin, Gong, et al. (2016) who used the asymptotic correlations of the overall and simple effect Cox model 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. By using the asymptotic correlations of the test statistics, they made a Dunnett-type correction which results in critical values that are less strict than if a Bonferroni correction were used.

In addition to the prevention and behavioral settings, factorial trials have also been widely used in cancer treatment trials. Freidlin and Korn (2017) conducted a survey of 30 two-by-two factorial cancer treatment trials published between 2007 and 2016. Such trials typically have a time-to-disease progression and/or death endpoint. They argue that in most cancer treatment settings, the no-interaction assumption cannot be justified. This is sometimes due to the overlapping toxicities of different treatments. Consequently, they discourage testing the overall effects in such a setting, and provide useful designs for testing the various simple effects.

While we will be focused on the time-to-event setting, there is no additional complexity to analyzing a continuous endpoint. For example, in the behavioral interventions setting, the Trials of Hypertension Prevention, Phase II (TOHP-II), used a two-by-two factorial design to study the effects of weight loss and sodium reduction on the change in diastolic blood pressure from baseline to termination visits (TOHP Collaborative Research Group, 1997). The change in systolic blood pressure and incidence of hypertension were important secondary endpoints. For the primary analysis, they used a fixed effects analysis of variance with terms for the weight loss, sodium reduction, and weight loss + sodium reduction groups. In a similar fashion, a dichotomous endpoint could be analyzed using a logistic regression model. All three types of endpoints (time to event, continuous, dichotomous) use analysis methods which rely on the asymptotic normality of the effect estimates.

The remainder of this paper is organized as follows. In Section 2, we establish notation and 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 derive formulas for asymptotic power and sample size. In Section 4, we use hypothetical, but realistic clinical trial scenarios to investigate the loss in power for the overall effects due to a subadditive interaction. In Section 5 we define and examine the power of three multiple testing procedures we propose for the overall and simple effects. Factorial trials, such as ACCORD BP, often consider the evaluations of A and B unrelated so that a two-sided 0.05 significance level is allocated to testing each of the overall A and B effects. We note that there remains debate on this issue (Byar and Piantadosi, 1985; Green, Liu, and O'Sullivan, 2002; Freidlin, Korn, and Gray, 2008). We assume such unrelatedness exists so our multiple testing procedures strongly, but separately, control the familywise type I error (FWE) for each family of hypotheses involving A, respectively, B, at the two-sided 0.05 significance level. In Section 6, we reanalyze the ACCORD-BP data with the proposed testing procedures. Both the power calculations and analysis use the R package `factorial2x2` (Leifer and Troendle, 2020). These calculations are demonstrated in the Appendix. Finally, in Section 7, we provide further discussion and recommendations.

2 Notation and Problem

We follow the set-up from Slud (1994). Consider a two-by-two 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 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 censoring time V_i which is conditionally independent of X_i given Z_i. We note from (4), there is no interaction when $\beta_3 = 0$, i.e., $\exp(\beta_3) = 1$, so the A and B effects are additive on the log hazard scale

$$\log(HR_A) + \log(HR_B) = \beta_1 + \beta_2 = \beta_1 + \beta_2 + \beta_3 = \log(HR_{AB})$$

which is equivalent to the A and B effects being multiplicative on the hazard scale

$$HR_A \cdot HR_B = \exp(\beta_1 + \beta_2) = \exp(\beta_1 + \beta_2 + \beta_3) = HR_{AB} \tag{5}$$

In contrast, since treatment benefit corresponds to a negative log hazard ratio, there is a subadditve interaction on the log scale when $\beta_3 > 0$ since then

$$\log(HR_A) + \log(HR_B) = \beta_1 + \beta_2 < \beta_1 + \beta_2 + \beta_3 = \log(HR_{AB})$$

or, equivalently,

$$HR_A \cdot HR_B = \exp(\beta_1) \cdot \exp(\beta_2) < \exp(\beta_1 + \beta_2 + \beta_3) = HR_{AB} \tag{6}$$

We quantify the magnitude of the interaction using its hazard ratio

$$HR_{int} = \exp(\beta_3) = \frac{HR_{AB}}{HR_A \cdot HR_B} \tag{7}$$

so a subadditive interaction corresponds to $HR_{int} > 1$. By similar reasoning, a superadditive interaction corresponds to $\beta_3 < 0$ or, equivalently, $HR_{int} < 1$.

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

$$H_{0A}^{\text{overall}} : \beta_1 = \beta_3 = 0 \qquad \text{and} \qquad H_{0B}^{\text{overall}} : \beta_2 = \beta_3 = 0 \tag{8}$$

As Slud (1994) points out, a factorial trial is ordinarily conducted to test whether either treatment A or B is beneficial in which case the appropriate null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. However, we choose to split H_0 into H_{0A}^{overall} and H_{0B}^{overall} for two reasons. First, as discussed for equation (1), since we have 50:50

randomization, $\log HR_A^{\text{overall}}$ is the average of the log hazard ratio of the A group to the C group, which is β_1, and the average of the log hazard ratio of the AB group to the B group, which is $\beta_1+\beta_3$. Thus, $\log HR_A$ is $\beta_1+0.5\beta_3$, so β_2 is not involved. Similarly, $\log HR_B = \beta_2 + 0.5\,\beta_3$, with no involvement of β_1. Second, in Section 5, the overall A and B effects will be separately tested as members of two different families of hypotheses. Writing their null hypotheses as in (8) will make this clearer. As Slud (1994) discusses, when the interaction β_3 is weak, which makes the overall A effect interpretable, a good choice to test H_{0A}^{overall} is 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

$$H_{0A}^{\text{simple}}:\beta_1=0,\qquad H_{0B}^{\text{simple}}:\beta_2=0,\qquad H_{0,AB}^{\text{simple}}:\beta_1+\beta_2+\beta_3=0 \tag{9}$$

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 for testing the overall effects; power for the simple effects

In this section, we focus on testing the overall A effect null hypothesis H_{0A}^{overall} in (8); similar calculations hold for testing the overall B effect. Since we use the stratified logrank statistic S_A^{str} to test H_{0A}^{overall}, we first review Slud's (1994) asymptotic mean formula for S_A^{str} to calculate power. This will also be used for the power studies presented in Sections 4 and 5.2

As discussed in the Appendix, since we assume that $\dfrac{n}{4}$ participants are randomized to each of the four groups C, A, B, and AB, Slud's formula for the asymptotic mean of the normalized stratified logrank statistic S_A^{str} for the overall A effect is

$$\mu_A^{str} \equiv \text{asympotic mean of } S_A^{str} = (\beta_1+0.5\beta_3)\sqrt{0.25\cdot n\cdot \Pr\{event\}} \tag{10}$$

In (10), the $\beta_1 + 0.5\beta_3$ term was discussed in the previous section while the probability of an event is averaged over the four groups:

$$\Pr\{event\} = \frac{1}{4}\big[\Pr\{event|C\} + \Pr\{event|A\} + \Pr\{event|B\} + \Pr\{event|AB\}\big] \tag{11}$$

If we want to test H_{0A}^{overall} at the two-sided α level, then to declare A efficacious, we need the test statistic $S_A^{\text{str}} < -z_{\alpha/2}$ where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. In the equation (4) parametrization of the Cox model, efficacy corresponds to rejecting H_{0A}^{overall} in favor of $H_{1A}^{\text{overall}} : \beta_1 + 0.5\beta_3 < 0$ (Slud, 1994). We note that the interaction β_3 appears in both the asymptotic mean μ_A^{str} and the alternative hypothesis H_{1A}^{overall}. However, unless we can statistically detect an interaction, i.e., infer that $\beta_3 \neq 0$, at the 0.05 significance level, rejecting H_{0A}^{overall} corresponds to declaring A efficacious. On the other hand, if a test of the interaction is significant at the 0.05 level, then we should declare A efficacious only if we reject H_{0A}^{simple} in favor of efficacy.

The usual normal theory calculations show that S_A^{str} has power to reject H_{0A}^{overall} in favor of H_{1A}^{overall} at the two-sided α significance level equal to

$$\text{power to reject } H_{0A}^{\text{overall}} = \Pr\{S_A^{str} < -z_{\alpha/2}\} = \Phi\Big(-z_{\alpha/2} - (\beta_1 + 0.5\beta_3)\sqrt{0.25 \cdot n \cdot \Pr\{event\}}\Big) \tag{12}$$

where $\Phi(\cdot)$ 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 (12) that power for the overall A effect diminishes as the size of the interaction, β_3, increases. So for a given $\beta_1 = \log(HR_A)$, there is less, respectively, greater, power for the overall A effect when there is a subadditve ($\beta_3 > 0$), respectively, superadditive ($\beta_3 < 0$), interaction as compared to no interaction ($\beta_3 = 0$). The number of participants n needed to have power $1 - \beta$ to reject H_{0A}^{overall} is given by

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2}{(\beta_1 + 0.5\beta_3)^2 \Pr\{event\}} \tag{13}$$

Similarly, the power and sample size to reject H_{0B}^{overall} is obtained by replacing β_1 with β_2 in (12) and (13).

In the Appendix, we review the power calculations for the simple effects' logrank statistics. In particular, the usual normal theory calculations show that S_A has power to reject H_{0A}^{simple} in favor of efficacy at the two-sided α significance level equal to

$$\text{power to reject } H_{0A}^{\text{simple}} = \Pr\{S_A < -z_{\alpha/2}\} = \Phi\Big(-z_{\alpha/2} - \beta_1\sqrt{0.25 \cdot \frac{n}{2} \cdot \Pr\{event|A \cup C\}}\Big) \tag{14}$$

We see in (14) that the power to reject H_{0A}^{simple} does not depend on the size of the interaction β_3. This is because the H_{0A}^{simple} compares the A group to the C group, without any involvement of treatment B. Similarly, the power to reject H_{0B}^{simple} does not depend on the interaction. However, the power to reject $H_{0,AB}^{\text{simple}}$ does depend on the interaction.

Lin, Gong, et al. (2016) show that under the null hypotheses H_{0A}^{simple}, H_{0B}^{simple}, and $H_{0,AB}^{\text{simple}}$, the stratified logrank statistic 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. This will be used for the Dunnett-type corrections for the multiple testing procedures to be introduced in Section 5.2.

4 Loss in power for the overall effects due to an interaction

In this section, we demonstrate how the power for the overall A effect decreases as the size of the interaction increases. We also compare the power for the overall A effect to the power for the simple A effect, which is unaffected by an interaction. We do this first by a theoretical calculation and then through several hypothetical, but realistic, trial scenarios. In particular, in Table 1 we consider seven scenarios to investigate power when there is no interaction (scenarios 1-2), a subadditive interaction (scenarios 3-6), and a superadditive interaction (scenario 7).

For our theoretical calculation, we recall that the power for the overall A effect is given by (12) and we see that the power decreases as the size of the interaction $\beta_3 = \log(HR_{int})$ increases. In contrast, β_3 does not appear in the power for the simple A effect as given by (14), so that power is unaffected by an interaction. To see when the simple effect's power becomes greater than the overall effect's power, we can set equal the two power formulas (12) and (14) and solve for β_3. To simplify this calculation, we note that for the realistic treatment effects we are considering, $\Pr\{event\}$ and $\Pr\{event | A \cup C\}$ are close enough to consider them to be equal. By straightforward algebra, we find that

$$\text{simple } A \text{ effect's power} > \text{overall } A \text{ effect's power} \quad \text{if and only if} \quad \beta_3 > (\sqrt{2} - 2)\,\beta_1 \tag{15}$$

To give a numerical example to gain insight into (15), suppose we assumed that the simple A effect $HR_A =$ 0.80. Then the simple A effect's power would be greater than the overall A effect's power if and only if

$HR_{int} = \exp(\beta_3) > \exp[(\sqrt{2} - 2)\log(0.8)] = 1.14$. To give $HR_{int} = 1.14$ further context, if we also assumed the same simple effect for B as for A, i.e., $HR_B = 0.80$, then the simple A effect's power would be greater than the overall A effect's power if and only if the simple AB effect $HR_{AB} = HR_A \cdot HR_B \cdot HR_{int} > 0.8 \cdot 0.8 \cdot 1.14 = 0.73$.

For for the clinical trial scenarios we now explore, we use some of the design parameters which were used for the ACCORD BP trial. In particular, we assume a 4.45% annual event rate in the control group and independent Uniform(4.0, 8.4) years censoring. We use a sample size of 4600 participants which has over 90% power to detect efficacy for each of the overall A and B effects assuming that:

1. A two-sided 0.05 significance level test will be used to separately test H_{0A}^{overall} and H_{0B}^{overall}.

2. The simple A and B effects each have hazard ratios $HR_A = HR_B = 0.80$.

3. There is no interaction so that (5) holds.

In addition to reporting the respective powers for the overall A and simple A effects in Table 1, we also include the power for the simple AB effect. Each of the three effects is tested at a two-sided Bonferroni-corrected 0.05/3 significance level.

In Table 1, we focus on how the size of the interaction affects the power to detect the overall A effect. The power for the overall A effect is the highest for scenario 1 where there is no interaction (i.e., interaction HR = 1) and we have the originally powered $HR_A = 0.80$. Scenario 2 also has no interaction, but has lower power due to the more modest risk reduction $HR_A = 0.85$. Scenario 3 has the originally powered $HR_A = 0.80$ with a small subadditive interaction, i.e., $HR_{int} = 1.08 > 1$. Thus the overall A effect's power is between that of scenarios 1 and 2. Scenario 4-6 also have the originally powered $HR_A = 0.80$, but with HR_{int} increasing from 1.13 to 1.33 so the power for the overall A effect is decreasing. Finally, scenario 7 has a superadditive interaction since $HR_{int} = 0.94 < 1$. Thus, although $HR_A = 0.90$ is less extreme than in scenario 6, the power for the overall A effect is higher than in scenario 6. As pointed out above, the simple A effect does not depend on HR_{int} so it has the same power in scenarios 1 and 3-6, all of which have $HR_A = 0.80$.

While our focus in Table 1 has been on the overall A effect's power, we note that across the scenarios, some combination of the overall A, simple A, and simple AB effects have reasonable power. This observation motivates the multiple testing procedures described in the next section.

5 Multiple testing procedures

5.1 Description of multiple testing procedures

In this section, we propose three multiple testing procedures for the A-related null hypotheses $\{H_{0A}^{\text{overall}}, H_{0A}^{\text{simple}}\}$ combined with $H_{0,AB}^{\text{simple}}$. We refer to the family of three null hypotheses $\{H_{0A}^{\text{overall}}, H_{0A}^{\text{simple}}, H_{0,AB}^{\text{simple}}\}$ as *Family 1.* Similar multiple testing procedures will be used for the B-related null hypotheses $\{H_{0B}^{\text{overall}}, H_{0B}^{\text{simple}}\}$ combined with $H_{0,AB}^{\text{simple}}$. We refer to the family of three null hypotheses $\{H_{0B}^{\text{overall}}, H_{0B}^{\text{simple}}, H_{0,AB}^{\text{simple}}\}$ as *Family 2.* We note that the intersection of Family 1 and Family 2 is $H_{0,AB}^{\text{simple}}$; this will be discussed further below.

The three multiple testing procedures for Family 1 correspond to three different ways to allocate the type I error among its three null hypotheses. The allocation will depend on the particular null hypotheses of interest, which will be influenced by our pre-trial belief about whether an interaction exists. Thus, we may be interested in all three null hypotheses, or, perhaps, a subset of the null hypotheses. We require that each of the proposed procedures strongly control the family wise error (FWE) at the two-sided 0.05 significance level for the null hypotheses being tested, whether it is all three hypotheses or a subset of the hypotheses. In a similar manner, Lin, Gong, et al. (2016) discussed the possibility of allocating the FWE among the three null hypotheses in various ways. Similarly, we strongly control Family 2's FWE at the two-sided 0.05 significance level.

With regard to allocating a two-sided 0.05 FWE to each of Family 1 and Family 2, this is because we view the the A and B research questions as being unrelated. Thus, if we were performing the standard factorial analysis, we would test each of the overall A and B effects at the 0.05 level. Under this perspective, we would separately control at the 0.05 level the FWE for the A-related null hypotheses $\{H_{0A}^{\text{overall}}, H_{0A}^{\text{simple}}\}$ and the FWE for the B-related null hypotheses $\{H_{0B}^{\text{overall}}, H_{0B}^{\text{simple}}\}$. However, we also want to test $H_{0,AB}^{\text{simple}}$, but without using any more type I error than is being allocated for the A-related and B-related null hypotheses. Thus, we include $H_{0,AB}^{\text{simple}}$ with both the A-related null hypotheses, to form Family 1, as well as the B-related null hypotheses, to form Family 2. In practice, we will apply the same (of the three possible) multiple testing procedure to Family 1 and Family 2. When there is no covariate adjustment in the logrank test statistics, we will see that the same nominal significance level will be used for testing $H_{0,AB}^{\text{simple}}$ as part of either Family 1 or Family 2. This means that $H_{0,AB}^{\text{simple}}$ is only tested once. When there is covariate adjustment, we will see that the significance level for $H_{0,AB}^{\text{simple}}$ can slightly differ between Family 1 and Family 2. This is due to the slightly different correlations between the covariate-adjusted logrank statistics as compared to the unadjusted logrank statistics. In this case, we can insure that $H_{0,AB}^{\text{simple}}$ is only tested once by testing it as part of Family 1, but not Family 2. It is important to emphasize that although $H_{0,AB}^{\text{simple}}$ is included in both Family 1 and Family 2 so as to control each family's FWE, it is a stand-alone hypothesis. Rejecting $H_{0,AB}^{\text{simple}}$ says nothing about the efficacy of A or B. It only says that their combination is effective.

We now describe the three new multiple testing procedures for Family 1. The same procedures are analogously defined for Family 2 with the intention that whichever of the procedures is chosen to be used for Family 1, it would also be used for Family 2. For the first new procedure, suppose we are equally interested in all three Family 1 null hypotheses and have no pre-trial knowledge about the presence (if any) of an interaction. Then we propose the *Equal Allocation 3* procedure which jointly tests H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$ using a common critical value. To calculate the critical value, we use a Dunnett-type multiple testing correction which depends on the $1/\sqrt{2}$ asymptotic correlation between the overall A and simple A (respectively, AB) logrank statistics, S_A^{str} and S_A (respectively, S_{AB}), as well as the 1/2 asymptotic correlation between the simple A and simple AB logrank statistics, S_A and S_{AB} (Slud 1994). In the Appendix, we show that the common critical value for testing is 2.32 and strongly controls the FWE at the 0.05 level. We note that the Dunnett-corrected 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.

For the second new procedure, if we are reasonably confident that no interaction exists, then H_{0A}^{overall} is the null hypothesis of primary interest. However, in the event that there were a small-to-moderate interaction, then $H_{0,AB}^{\text{simple}}$ would become the hypothesis of interest. This is because the simple AB effect could be more beneficial than the simple A effect such as in scenario 4 from Table 1. In this case, we jointly test only H_{0A}^{overall} and $H_{0,AB}^{\text{simple}}$, but do not test H_{0A}^{simple}, using the *Proportional Allocation 2* procedure. The Proportional Allocation 2 procedure "spends" most its error testing H_{0A}^{overall}. In particular, H_{0A}^{overall} is tested at the two-sided $(2/3) \cdot 0.05 = 0.033$ level corresponding to a critical value of 2.13. We then use a Dunnett-type correction and the $1/\sqrt{2}$ asymptotic correlation between the overall A and simple AB logrank statistics to obtain a critical value of 2.24 for $H_{0,AB}^{\text{simple}}$. This strongly controls the FWE for H_{0A}^{overall} and $H_{0,AB}^{\text{simple}}$ at the two-sided 0.05 level. We note that the Dunnett-corrected 2.24 corresponds to a two-sided 0.0251 significance level to test $H_{0,AB}^{\text{simple}}$ which is larger than a Bonferroni-corrected $0.05/3 = 0.0167$ significance level.

For the third new procedure, if we are reasonably confident that an interaction exists, then we do not test H_{0A}^{overall}. Instead, we propose the *Equal Allocation 2* procedure which jointly tests H_{0A}^{simple} and $H_{0,AB}^{\text{simple}}$ using the same critical value. To calculate that critical value, we use a Dunnett-type correction and the 1/2 asymptotic correlation between the simple A and simple AB logrank statistics to calculate the common critical value = 2.22. This strongly controls the FWE for H_{0A}^{simple} and $H_{0,AB}^{\text{simple}}$at the two-sided 0.05 level. We note that the Dunnett-corrected 2.22 corresponds to a two-sided significance level of 0.0264 which is slightly larger than the Bonferroni-corrected $0.05/2 = 0.025$ two-sided significance level.

In the Appendix, we show how to use the R package `factorial2x2` (Leifer and Troendle, 2020) to calculate the power for a trial as well as analyze a trial using the Equal Allocation 3 procedure. The `factorial2x2` package can be similarly used for the Proportional Allocation 2 and Equal Allocation 2 procedures. We also provide in the Appendix a small simulation study which empirically demonstrates the FWE control of the three procedures when there are adjustment covariates. This is important since in an actual analysis, the correlations need to be estimated from the data.

5.2 Power comparisons for the multiple testing procedures

In Table 2, we investigate the power of the three multiple testing procedures using the Table 1 scenarios. For comparative purposes, we also include the power detect the overall A and B, each tested at the two-sided 0.05 significance level. We call this test of an overall effect the *Factorial* procedure.

For each of the scenarios, we put in boldface the effect and corresponding power we most want to detect. For scenario 1, treatment A and the combination treatment AB have the same efficacy $HR = 0.80$, but we take the perspective that one treatment is better than two to achieve the same effect. In this case, we would be most interested in declaring A efficacious either through the overall A effect or the simple A effect.

Thus, the power for *any* A effect is most relevant since that is the power that H_{0A}^{overall} or H_{0A}^{simple} is rejected in favor of efficacy. Since there is no interaction, the Factorial procedure has the highest power, followed closely by the Proportional Allocation 2 and Equal Allocation 3 procedures since both of those procedures also test the overall A effect. The Equal Allocation 2 procedure has noticeably lower power since it does not test the overall A effect. The situation is similar for scenario 2 which also has no interaction as well as scenario 3 which has a small interaction. We note that all procedures appropriately control the *any* B effect at the one-sided 2.5% level since B has no effect in scenarios 1 and 2, and a harmful effect in scenario 3.

In scenario 4, the combination AB is the most efficacious, and all three multiple comparison procedures have over 90% power to detect it. Of course the Factorial procedure cannot detect the simple AB effect since it does not test for it. Scenarios 5 and 6 have moderate-to-large interactions so the Factorial and Proportional Allocation 2 procedures have noticeably lower power to detect the *any* A, respectively, *any* B, effect than the Equal Allocation 3 and Equal Allocation 2 procedures. This is because the latter two procedures test the simple A and simple B effects while the former two procedures do not. For scenario 7's superadditive interaction, all three multiple comparison procedures have over 90% power to detect the AB combination, while the Factorial procedure does not test for it. Looking across the scenarios in Table 2, the general finding is that the Equal Allocation 3 procedure is the most robust for declaring efficacious the effect of greatest interest.

6 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 (A) control and intensive glcymeia (B) 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 Equal Allocation 3, Proportional Allocation 2, and Equal Allocation 2 procedures are slightly different from those reported in Section 5.1. This is because the Section 5.1 significance levels depend on the asymptotic correlations between the *unadjusted* logrank statistics for the overall A, simple A, and simple AB effects. 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, 2020) to calculate the correlation matrices to be:

$$\begin{array}{r} \text{overall } A \\ \text{simple } A \\ \text{simple } AB \end{array} \begin{pmatrix} 1 & 0.736 & 0.731 \\ & 1 & 0.427 \\ & & 1 \end{pmatrix} \qquad \begin{array}{r} \text{overall } B \\ \text{simple } B \\ \text{simple } AB \end{array} \begin{pmatrix} 1 & 0.724 & 0.735 \\ & 1 & 0.421 \\ & & 1 \end{pmatrix} \tag{16}$$

We see for the left-hand matrix that the correlation between the covariate-adjusted overall and simple effect estimates, 0.736 and 0.731, 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, 0.427, is smaller than the asymptotic correlation $1/2 = 0.5$ between the unadjusted simple estimates. Similarly for the right-hand matrix. Using the left-hand (respectively, right-hand) matrix, `fac2x2analyze` calculates that the Equal Allocation 3 procedure tests each of the Family 1 (respectively, Family 2) null hypotheses at the two-sided 0.0208 level; the Proportional Allocation 2 procedure tests the overall A (respectively, B) hypothesis at the 0.033 level and the simple AB hypothesis at the 0.026 level; and the Equal Allocation 2 procedure tests the simple A (respectively, B) and simple AB hypotheses at the 0.026 level. As seen in Table 3, the overall glycemia effect is significant with respect to the Proportional Allocation 2 procedure, the simple glycemia effect is significant with respect to the Equal Allocation 3 and Equal Allocation 2 procedures, and the simple BP + glycemia effect is significant with respect to all three procedures. Based on the Table 3 results for the ACCORD BP trial, intensive glycemia control alone improved the primary endpoint of major CVD outcome without additional benefit from combining it with strict blood pressure control.

7 Discussion

The two-by-two factorial trial design provides the opportunity to compare each of two treatments A and B, as well as their combination AB, to a control group C. When there is good reason to believe that no interaction exists between A and B, such as may be the case when treatment-associated toxicities are low, hypothesis tests of the overall A and B effects are more appropriate than tests of the simple A and B effects. This is because when there is no interaction, the overall A and B effects equal the simple A and B effects, respectively, but the overall effects are each tested using all trial participants while the the simple effects are tested using half of the trial participants. Hence the overall effects' tests are more powerful. Moreover, a test of the simple AB effect can be powerful since no interaction is equivalent to the simple A and B effects being additive on an appropriate scale, such as the log hazard scale for the Cox model. On the other hand, if we believe an interaction exists, only the simple effects should be tested since the overall effects have lower power and are not equal to the simple effects. Indeed, when there is an interaction, the simple, not the overall, A and B effects are the appropriate measures of treatment A and B efficacy and the power to detect the simple effects is unaffected by an interaction.

This paper is interested in the situation when we are unsure about whether an interaction exists in the time-to-event endpoint setting. We wanted to gain a better understanding about how to jointly test the overall and/or simple effects to have robust power for detecting the most efficacious treatment. Through hypothetical, but realistic clinical trial scenarios, we first investigated the relationship between the interaction hazard ratio and the loss in power for the overall effects. For our power studies, we used Slud's (1994) and Schoenfeld's (1981) asymptotic results for the Cox proportional hazards model with independent censoring. As demonstrated by Wittes and Brittain (1989) for normally distributed outcomes, we found that a moderate interaction HR compared to the control group causes a dramatic diminution in the power to detect the overall effects. Nevertheless, sometimes there can be reasonable power to detect the simple AB effect when it has a more extreme HR than either the simple A or simple B effects.

We then proposed three multiple testing procedures for the A-related null hypotheses $\{H_{0A}^{\text{overall}}, H_{0A}^{\text{simple}}\}$

as well as $H_{0,AB}^{\text{simple}}$, the combination of which we called Family 1, to provide robust power to detect the most efficacious treatment. Similar joint testing would be carried out for the B-related null hypotheses $\{H_{0B}^{\text{overall}}, H_{0B}^{\text{simple}}\}$ which in combination with $H_{0,AB}^{\text{simple}}$ we called Family 2. Throughout this paper, we assumed that the A and B research questions were unrelated so that joint testing of the Family 1, respectively, Family 2, null hypotheses would each be done to preserve the FWE at the 0.05 level for each family separately. All three procedures use Dunnett-type corrections to calculate the critical values for significance by using the asymptotic correlations between the logrank test statistics used to test the null hypotheses. For a situation in which we had no pre-trial knowledge about the existence of an interaction, we proposed the Equal Allocation 3 procedure which equally allocates the 0.05 type I error across each family's three null hypotheses. When we were fairly certain that little or no interaction existed, we proposed the Proportional Allocation 2 procedure which allocates 2/3 of the type I error to test the overall A (respectively B) effect, and the remaining Dunnett-corrected type I error to test the simple AB effect. Finally, when we were fairly certain that an interaction existed, we proposed the Equal Allocation 2 procedure which does not test the overall effect, but equally allocates the type I error to test the simple A (respectively, B) effect and the simple AB effects.

We next investigated the power of the three testing procedures using the same hypothetical clinical trial scenarios we used for studying the interaction-related loss in power for the overall effects. Across the scenarios we considered, the Equal Allocation 3 procedure had the most robust power for declaring the most efficacious treatment significantly better than control. When there was little or no interaction, the Proportional Allocation 2 procedure had slightly better power than the Equal Allocation 3 procedure and both had noticeably better power than the Equal Allocation 2 procedure. This is because the Proportional Allocation 2 and Equal Allocation 3 procedures test the overall effects which have higher power when the interaction is weak than the simple effects, while the Equal Allocation 2 procedure does not test the overall effects. On the other hand, when there is a moderate interaction, the Equal Allocation 2 procedure had slightly higher power than the Equal Allocation 3 procedure and sometimes noticeably higher power than the Proportional Allocation 2 procedure. This is because a moderate interaction causes a reduction in power for the overall effect which is appropriate since the overall effect is different from the simple effect and is thus hard to interpret. In this case, hypothesis testing should concentrate on the simple A, B, and AB effects, which are tested by the Equal Allocation 2 and Equal Allocation 3 procedures, but not the Proportional Allocation 2 procedure which only tests the simple AB effect in addition to the overall effects.

A separate issue which we did not investigate is whether the A and B research questions should be considered unrelated, as we have assumed, or related with respect to controlling type I error. This remains a question of debate. When the research questions are considered related, Korn and Freidlin (2016) have proposed a three-step procedure with sequentially ordered simple effects' hypotheses which controls the familywise type I error across all of the A and B comparisons. Korn and Freidlin assume that an interaction is likely, such as in the cancer setting in which treatments can have associated toxicities, so they only consider tests involving the simple A, B, and AB effects, but not the overall effects. For the Table 2 sample size assumptions on the control group event rate, independent censoring distribution, $HR_A = HR_B = 0.80$, and no interaction so $HR_{AB} = HR_A \times HR_B = 0.64$, their procedure requires about 64% more participants than the 4600 participants for Table 2. The first step of their procedure tests each of the simple effects at the two-sided 0.05/3 significance level and has 80% power to detect each of the simple A and B effects, and over 99% power to detect the simple AB effect. 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 all of the simple effects' null hypotheses, an interaction may exist, comparisons between the simple effects are desired, 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 unrelated, and the extent to which an interaction may exist. When the research questions are unrelated, the Equal Allocation 3 procedure can be useful as it exploits the sample size advantage of the overall A and B tests when there is no interaction, but also tests the simple A and B effects, which are unaffected when an interaction exists, as well as the simple AB effect.

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. Byar DP and Piantadosi S. Factorial designs for randomized clinical trials. *Cancer Treat Rep.* 1985; 69(10):1055-1063.

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

4. 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.

5. 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.

6. Freidlin B, Korn El, Gray R, et al. Multi-arm clinical trials of new agents: Some design considerations. *Clin Cancer Res.* 2008; 14(14): 4368-4371.

7. Freidlin B and Korn EL. Two-by-two factorial cancer treatment trials: is sufficient attention being paid to possible interactions? *J Natl Cancer Inst.* 2017;109(9).

8. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-0. 2020. https://CRAN.R-project.org/package=mvtnorm.

9. 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.

10. Green S, Liu PY, and O'Sullivan J. Factorial design considerations. *J clin Oncol.* 2002;20(16):3424-3430,

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

12. Leifer ES and Troendle JF. factorial2x2: Design and Analysis of a 2x2 factorial trial. R package version 0.1. 2020. https://CRAN.R-project.org/package=factorial2x2.

13. 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.

14. 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.

15. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 $\times k$ factorial design when time-to-failure is the outcome. *Control Clin Trials.* 1993; 14:511-522.

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

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

18. The Trials of Hypertension Prevention Collaborative Research Group. The Trials of Hypertension Prevention, phase II. *Arch Intern Med.* 1997;157:657-67.

A Appendix

A.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

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

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 (10), 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

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

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

$$\mu_A = \beta_1 \sqrt{0.25 \cdot \frac{n}{2} \cdot \Pr\{event | A \cup C\}} \tag{A.3}$$

The usual normal theory calculations show that S_A has power to reject H_{0A}^{simple} in favor of efficacy at the two-sided α significance level equal to

$$\text{power to reject } H_{0A}^{\text{simple}} = \Pr\{S_A < -z_{\alpha/2}\} = \Phi\left(-z_{\alpha/2} - \beta_1 \sqrt{0.25 \cdot \frac{n}{2} \cdot \Pr\{event | A \cup C\}} \right) \tag{A.4}$$

A.2 The Equal Allocation 3 procedure: strong FWE control, power calculations, and analysis using the `factorial2x2` R package

We demonstrate how to design and analyze a two-by-two factorial trial using the `factorial2x2` R package (Leifer and Troendle, 2020). Our demonstration uses the Equal Allocation 3 procedure, but the Proportional

Allocation 2 and Equal Allocation 2 procedures could be used in a similar manner. We focus on the Family 1 null hypotheses H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$, but the same remarks apply to the Family 2 null hypotheses. First we calculate the 0.0203 two-sided significance level, which corresponds to a 2.32 critical value, which the Equal Allocation 3 procedure uses to test each of the Family 1 null hypotheses. Then we show how the 0.0203 significance level strongly controls the FWE at the two-sided 0.05 significance level. We next show how to power a trial using the Table 2 design parameters. Finally, we analyze simulated trial data which are available with the `factorial2x2` package.

The Equal Allocation 3 procedure tests H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$ using the stratified logrank statistic S_A^{str} and the simple logrank statistics S_A and S_{AB}, respectively. We reject the respective null hypotheses for large absolute values of the corresponding statistics. Assuming all three null hypotheses are true, then by Slud (1994), $(S_A^{\text{str}}, S_A, S_{AB})$ are jointly asymptotically normal with mean vector $(0,0,0)$ and correlation matrix (approximating $1/\sqrt{2}$ by 0.707)

$$\begin{array}{c} S_A^{\text{str}} \\ S_A \\ S_{AB} \end{array} \begin{pmatrix} 1 & 0.707 & 0.707 \\ & 1 & 0.5 \\ & & 1 \end{pmatrix} \tag{A.5}$$

The critical value 2.32 is the solution to c of

$$0.05 = 1 - \Pr\{|S_A^{\text{str}}| < c, |S_A| < c, |S_{AB}| < c\} \tag{A.6}$$

The `crit2x2` function from `factorial2x2` R package uses the `mvtnorm` R package (Genz, Bretz, et al. 2020) to calculate $c = 2.32$:

```
crit2x2(corAa = 0.707, corAab = 0.707, coraab = 0.5, alpha = 0.05)
$critEA3
[1] 2.32          # common critical value
$sigEA3
[1] 0.02034088   # common two-sided significance level
```

Next we show that the Equal Allocation 3 procedure strongly controls the FWE at the two-sided 0.05 level. To do this, we need to show that there is no more than a 0.05 probability of falsely rejecting at least one true null hypothesis among H_{0A}^{overall}, H_{0A}^{simple}, $H_{0,AB}^{\text{simple}}$. We just saw that taking $c = 2.32$ in (A.6), there is a 0.05 probability of falsely rejecting at least one of H_{0A}^{overall}, H_{0A}^{simple}, $H_{0,AB}^{\text{simple}}$ assuming all three null hypotheses are true. However, suppose only H_{0A}^{overall} and H_{0A}^{simple} are true. Then by Slud (1994), (S_A^{str}, S_A) is asymptotically normal where S_A^{str} and S_A each have mean 0, variance 1, and their correlation is $1/\sqrt{2}$. Then the probability of falsely rejecting at least one of H_{0A}^{overall} and H_{0A}^{simple} is less than 0.05 since (we use that 2.24 corresponds to a two-sided 0.025 significance level):

$$\begin{aligned}\Pr\{|S_A^{\text{str}}| > 2.32 \cup |S_A| > 2.32\} &< \Pr\{|S_A^{\text{str}}| > 2.24 \cup |S_A| > 2.24\} \\ &< \Pr\{|S_A^{\text{str}}| > 2.24\} + \Pr\{|S_A| > 2.24\} = 0.025 + 0.025 = 0.05\end{aligned}$$

We similarly argue for any other subset of true null hypotheses.

Next we show how to use the `factorial2x2` package to power a trial in which we intend to use the Equal Allocation 3 procedure. We assume that we are given the trial parameters from Table 2. These include the trial's expected accrual and follow-up period as well as the expected event rate for participants who are in the control C arm. In the Table 2 power studies, we assumed the ACCORD-BP trial's estimated 4.4 years of uniform accrual followed by 4 years of additional follow-up. We also assumed the expected 4.45% annual event rate in the control group. Next we assume a variety of hypothesized values for the simple A, B, and AB effects such as those in Table 2. For example, for scenario 4 of Table 2, we assumed the simple effect hazard ratios (as compared to the C group) $HR_A = 0.80$, $HR_B = 0.80$, and $HR_{AB} = 0.72$. So from (7) the interaction hazard ratio is

$$HR_{int} = \exp(\beta_3) = \frac{HR_{AB}}{HR_A \cdot HR_B} = \frac{0.72}{0.8 \cdot 0.8} = 1.125 \tag{A.7}$$

which is a subadditive interaction since $HR_{int} > 1$. With these parameters, we can use the `fac2x2design` function from the `factorial2x2` package to see that a sample size of 4600 participants provides 90.7% power for the Equal Allocation 3 procedure to detect the most efficacious treatment, which in this case is the AB combination:

```
n <- 4600           # total sample size
rateC <- 0.0445     # one year event rate in the control group
hrA <- 0.80         # simple A effect hazard ratio
hrB <- 0.80         # simple B effect hazard ratio
hrAB <- 0.72        # simple AB effect hazard ratio
mincens <- 4.0      # minimum censoring time in years
maxcens <- 8.4      # maximum censoring time in years
fac2x2design(n, rateC, hrA, hrB, hrAB, mincens, maxcens, alpha = 0.05)
$powerEA3overallA
```

```
 [1] 0.5861992      # power for the overall A effect
$powerEA3simpleA
 [1] 0.5817954      # power for the simple A effect
$powerEA3simpleAB
 [1] 0.9071236      # power for the simple AB effect
$powerEA3$anyA
 [1] 0.7060777      # power for overall or simple A effects
```

Since $HR_A = HR_B = 0.80$ in scenario 4, the above power calculations are the same for the Family 2 null hypotheses.

Next we demonstrate how `fac2x2design` computes the 70.6% power for the overall or simple A effects. First, `fac2x2design` computes the asymptotic mean vector (μ_A^{str}, μ_A). To compute μ_A^{str} from (10), we use $\beta_1 = \log(HR_A) = \log(0.80)$, $\beta_3 = \log\left(\frac{HR_{AB}}{HR_A \cdot HR_B}\right) = \log\left(\frac{0.72}{0.80 \cdot 0.80}\right) = \log(1.125)$, and $n = 4600$, so `fac2x2design` computes $\Pr\{event\} = 0.208$ and $\mu_A^{str} = -2.538$. To compute the asymptotic mean of μ_A from (A.3), we have $\beta_1 = \log(0.80)$ and $n = 2300$ so `fac2x2design` computes $\Pr\{event|A \cup C\} = 0.223$ and $\mu_A = -2.526$. Then using (μ_A^{str}, μ_A) and the correlation matrix (A.5), `fac2x2design` computes

$$\Pr\{[S_A^{\text{str}} < -2.32] \cup [S_A < -2.32]\} = 0.706. \tag{A.8}$$

It is important to note that under the local alternatives in this scenario, as opposed to the intersection of the null hypotheses $H_{0A}^{\text{overall}} \cap H_{0A}^{\text{simple}} \cap H_{0,AB}^{\text{simple}}$, the asymptotic correlation matrix is slightly different from (A.5). However, this difference 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.

Finally, we show how to use the `fac2x2analyze` function from the `factorial2x2` R package to analyze simulated trial data using the Equal Allocation 3 procedure. We use the simulated data in `simdata` which is a 4600-by-9 matrix which is loaded with `factorial2x2` so that the reader may reproduce the below results. `simdata` corresponds to a simulated 2x2 factorial clinical trial of 4600 participants which is similar to the ACCORD-BP trial. Participants are randomized in equal proportions to groups C, A, B, or AB. Baseline covariates include whether the participant has a history of cardiovascular disease (yes/no) and which clinical center randomized the participant. The true HRs vs. the control group C were simulated to correspond to scenario 6 in Table 2: $HR_A = 0.80$, $HR_B = 0.80$, and $HR_{AB} = 0.72$. As seen in (A.7), this corresponds to a true AB interaction of 1.125.

Prior to the trial, it has been decided that the A and B research questions are sufficiently unrelated so that we will use the Equal Allocation 3 procedure to have 0.05 level FWE control of Family 1, respectively, Family 2. Regarding Family 1, we see from the R code and output below that the estimated overall A effect HR is 0.89 (95% CI: $0.79 - 1.01$, $p = 0.061$), the estimated simple A effect HR is 0.81 (95% CI: $0.68 - 0.96$, $p = 0.015$), and the estimated simple AB effect HR is 0.76 (95% CI: $0.64 - 0.90$, $p = 0.002$). We also see that a 2.31 critical value, corresponding to a 0.0209 two-sided significance level, is used for testing each of the Family 1 null hypotheses. The 0.0209 significance level, which is adjusted for the baseline covariates, is very slightly different from the 0.0203 significance level when there is no covariate adjustment. There is more discussion on this point in Section 6. Thus, we may reject the null hypotheses H_{0A}^{simple} and $H_{0,AB}^{\text{simple}}$, but not H_{0A}^{overall}. We conclude that A and AB are efficacious treatments. We note that according to scenario 4 in Table 2 from which we simulated the data, Equal Allocation 3 provided respective pre-trial powers 58.6%, 58.2%, and 90.7% for rejecting H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$. Of course, without knowledge of the true data generating mechanism, those pre-trial powers would not be known to us. We also note that the estimated interaction HR is 1.23 (95% CI: $0.97 - 1.58$, $p = 0.093$). Thus, the true interaction, $HR_{int} = 1.125$, is not quite detected although there is a trend. Nevertheless, the overall A effect was not significant while the simple A effect was significant so we would prefer to measure treatment A efficacy through the simple A effect.

Regarding the Family 2 null hypotheses, we see from the R code and output below that the estimated overall B effect HR is 0.85 (95% CI: $0.75 - 0.96$, $p = 0.008$) and the estimated simple B effect HR is 0.77 (95% CI: $0.65 - 0.91$, $p = 0.002$). The estimated simple AB effect HR, 95% confidence interval, and p-value is as reported in the previous paragraph. We also see that a 2.32 critical value, corresponding to a 0.0203 two-sided significance level, is used for testing each of the three null hypotheses H_{0B}^{overall}, H_{0B}^{simple}, and $H_{0,AB}^{\text{simple}}$, although $H_{0,AB}^{\text{simple}}$ was already tested as part of Family 1. We see that all three null hypotheses are rejected by the Equal Allocation 3 procedure. We conclude that B and AB are efficacious treatments. As noted above, since the data were simulated according to scenario 4 in Table 2, and since the HRs in that scenario are symmetric in A and B, Equal Allocation 3 provided respective pre-trial powers of 58.6%, 58.2%, and 90.7%, for rejecting H_{0B}^{overall}, H_{0B}^{simple}, and $H_{0,AB}^{\text{simple}}$. We note that although H_{0A}^{overall} was not rejected, H_{0B}^{overall} was rejected. This was due to the randomness associated with using a single simulated trial. Nevertheless, since both H_{0B}^{overall} and H_{0B}^{simple} were rejected, we would have to decide on whether to use the estimated overall B effect HR = 0.85 or the estimated simple B effect HR = 0.77 as the measure of the treatment B effect. Recall that we used the estimated simple A effect to measure the treatment A effect due to its statistical significance, the non-significance of the estimated overall A effect, and the trend towards an A-by-B interaction. The trend towards an interaction which could be associated with the non-significance of the overall A effect makes us prefer the estimated simple B effect to the estimated overall B effect.

Now, A, B, and AB are all significantly efficacious as compared to control, but none of the three is superior to the other two. If we believe a single treatment is better than the combination when they are similarly efficacious, then we would recommend either A or B for treatment. Of course, we know that the true $HR_{AB} = 0.72 < 0.80 = HR_A = HR_B$ so AB is more efficacious than A or B. However, to have good power to establish the superiority of AB over A and B would require a substantially larger sample size using a sequential hypothesis testing procedure such as the one proposed by Korn and Freidlin (2016).

```
#### R code and output for analyzing the simulated trial data
time <- simdata[, ’time’]    # follow-up time
event <- simdata[, ’event’]  # event indicator
indA <- simdata[, ’indA’]    # treatment A indicator
indB <- simdata[, ’indB’]    # treatment B indicator
covmat <- simdata[, 6:10]
# simdata[, 6:10] corresponds to the baseline covariates which include
# a history of cardiovascular disease (yes/no) and four indicator
# variables (yes/no) which correspond to which of 5 clinical centers enrolled each
# of the participants
fac2x2analyze(time, event, indA, indB, covmat, alpha = 0.05)
$hrAoverall
[1] 0.8895135          # overall A effect HR
$ciAoverall
[1] 0.786823 1.005607  # 95% CI for overall A effect HR
$pvalAoverall
[1] 0.06139083         # p-value for overall A effect HR
$hrAsimple
[1] 0.8096082          # simple A effect HR
$ciAsimple
[1] 0.6832791 0.9592939 # 95% CI for simple A effect HR
$pvalAsimple
[1] 0.01468184         # p-value for simple A effect HR
$hrABsimple
[1] 0.7583061          # simple AB effect HR
$ciABsimple
[1] 0.6389355 0.8999785 # 95% CI for simple A effect HR
$pvalABsimple
[1] 0.001545967         # p-value for simple AB effect
$hrABint
```

```
[1] 1.234542            # interaction HR
$ciABint
[1] 0.9654633 1.5786129 # 95% CI for interaction HR
$pvalABint
0.09299803              # p-value for interaction HR
$critEA3_A
[1] 2.31                # Equal Allocation 3 procedure's critical value for Family 1's
                        # hypotheses tests
$sigE3_A
[1] 0.02091404         # Equal Allocation 3 procedure's significance level for rejection
                       # for Family 1's hypotheses tests
$resultEA3_A
[1] "accept overall A" "reject simple A" "reject simple AB"  # hypotheses tests' results
$hrBoverall
[1] 0.8466461             # overall B effect HR
$ciBoverall
[1] 0.7486842 0.9574258   # 95% CI for overall B effect HR
$pvalBoverall
[1] 0.007966693           # p-value for overall B effect HR
$hrBsimple
[1] 0.7654237             # simple B effect HR
$ciBsimple
[1] 0.6452766 0.9079416   # 95% CI for simple B effect HR
$pvalBsimple
[1] 0.002150925           # p-value for simple B effect HR
$critEA3_B
[1] 2.32                 # Equal Allocation 3 procedure's critical value for Family 2's
                          # hypotheses tests
$sigEA3_B
[1] 0.02034088            # Equal Allocation 3 procedure's significance level for rejection
                          # for Family 2's hypotheses tests
$resultEA3_B
[1] "reject overall B" "reject simple B"  "reject simple AB" # hypothesis tests' results
```

A.3 Simulation study demonstrating the multiple testing procedures' family-wise error control when adjusting for covariates

Here we report the results of a small simulation study we did to examine the 0.05 level FWE control of the three multiple testing procedures when adjusting for covariates. Our general finding is that the FWE is well controlled. In our simulation study, the FWE control is with respect to the Family 1 null hypotheses H_{0A}^{overall}, H_{0A}^{simple}, and $H_{0,AB}^{\text{simple}}$. However, similar results would apply to the Family 2 null hypotheses. Since we are checking the FWE, we assume that

$$\beta_1 = \beta_2 = \beta_3 = 0 \tag{A.9}$$

in the Cox model (4). This case has a higher FWE than when only one or two of the β's are zero. Thus, if FWE is controlled for (A.9), we can be confident that FWE will be controlled when only one or two of the

β's are zero.

First we explain how we conducted our simulation study. We ran ten simulations, each with $m = 10{,}000$ simulated trials. Each simulation had either $n = 500$ or $n = 1000$ participants. Those participants were obtained by sampling with replacement from the ACCORD BP trial data set. Associated with each of the $i = 1, 2, \ldots, n$ participants was a risk score r_i. The risk score was defined to be

$$r_i = \exp(\widehat{\beta}\, z_i) \tag{A.10}$$

In (A.10), z_i is participant i's vector of seven binary covariates used for the primary ACCORD BP analysis. Those covariates include an indicator of whether the participant had a previous cardiovascular event as well as six additional indicators specifying which of the seven clinical center networks enrolled the participant. In (A.10), $\widehat{\beta}$ is the partial maximum likelihood estimator obtained by fitting a Cox model to the ACCORD BP data which only included terms for the covariates. Then for each of the $j = 1, 2, \ldots, 10000$ trials within a simulation, we generated participant i's survival time X_{ij} according to the hazard function

$$\lambda_0(t|\alpha,\, \gamma,\, r_i) = \alpha t^{\gamma} r_i \tag{A.11}$$

Thus, participant i's event time was generated using a Weibull nuisance hazard function with shape parameter α and scale parameter γ with r_i entering as a multiplicative constant. In other words, we assumed a Cox model with Weibull nuisance hazard function with covariates z_i, but no A and B treatment effect. We also generated a censoring time V_{ij} according to an independent Uniform$(4, 8.4)$ years distribution, as was assumed for Tables 1 and 2 in our paper. Thus participant i's "observed" data in trial j was an observed follow-up time $T_{ij} = \min(X_{ij}, V_{ij})$ and an observed event indicator $\delta_{ij} = I(X_{ij} \le V_{ij})$.

The results of our simulation study are given in Table 4. We studied various levels of censoring as well as varying the scale and shape parameters of the Weibull nuisance hazard function. Our general finding is that the FWE for the three multiple testing procedures is quite close to the desired 0.05 level. Moreover, when it is slightly elevated above 0.05, it is because one of the logrank test statistics is slightly elevated above 0.05. To elaborate, since each simulation had 10,000 trials, the lower 95% confidence bound for each of the simulated rejection rates $\widehat{p}$ is approximately

$$\widehat{p} - 1.96\sqrt{\frac{\widehat{p}\cdot(1-\widehat{p})}{10000}} \tag{A.12}$$

Thus, without making any multiple comparisons correction for the 10 simulations × 6 rates per simulation = 60 reported rates in Table 4, any rate above 5.44% might be considered elevated beyond Monte Carlo error. Since we adjusted for seven binary covariates, we believe any slightly inflated type I errors for the logrank statistics indicate that the corresponding asymptotic results have not fully taken effect for the sample sizes and number of events we studied. Note that for the heaviest censoring cases in simulations 1 and 2, we only ran simulations for $n = 1000$ participants, but not $n = 500$ participants, to insure an adequate number of events in each of the four groups C, A, B, AB in the two-by-two factorial design while adjusting for seven binary covariates.

Table 1: Power for overall and simple effects: $\boldsymbol{n}$**=4600**[a]

Scenario	True HRs vs. C[b]				Power[c] (%) to declare a benefit for:		
	HR_A	HR_B	HR_{AB}	HR_{int}	overall A	simple A	simple AB
1	0.80	1.00	0.80	1.00	88.0	55.3	55.3
2	0.85	1.00	0.85	1.00	59.3	29.8	29.8
3	0.80	1.10	0.95	1.08	74.1	55.3	3.7
4	0.80	0.80	0.72	1.13	55.5	55.3	89.4
5	0.80	0.82	0.80	1.22	32.3	55.3	55.3
6	0.80	0.75	0.80	1.33	12.1	55.3	55.3
7	0.90	0.85	0.72	0.94	39.6	12.1	89.4

[a] True hazard function is given in equation (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 = 897-1087, depending on the scenario.

[b] HR_A, HR_B, HR_{AB} are the respective hazard ratios of the A, B, and AB groups vs. the control group. HR_{int} = interaction hazard ratio.

[c] Power for the overall A effect is computed using (12). Power for the simple A effect is computed using (14); a similar formula is used for the simple AB effect. All tests use a two-sided 0.05/3 significance level.

Table 2: Power for multiple testing procedures: $\boldsymbol{n}$=**4600**[a]

Scenario	True HRs vs. C				Power (%) to declare a statistically significant benefit for							
	HR_A	HR_B	HR_{AB}	HR_{int}	proc[b]	overall A	simple A	any A[c]	overall B	simple B	any B[c]	simple AB
1	**0.80**	1.00	0.80	1.00	EA3	89.5	58.2	**90.4**[d]	1.0	1.0	1.8	58.2
					PA2	92.5	X[e]	**92.5**	1.7	X	1.7	61.3
					EA2	X	62.0	**62.0**	X	1.3	1.3	62.0
					FAC	94.7	X	**94.7**	2.5	X	2.5	X
2	**0.85**	1.00	0.85	1.00	EA3	62.3	32.4	**65.0**	1.0	1.0	1.8	32.4
					PA2	69.3	X	**69.3**	1.7	X	1.7	35.3
					EA2	X	36.0	**36.0**	X	1.3	1.3	36.0
					FAC	75.0	X	**75.0**	2.5	X	2.5	X
3	**0.80**	1.10	0.95	1.08	EA3	76.7	58.2	**80.8**	< 0.1	< 0.1	< 0.1	4.3
					PA2	82.1	X	**82.1**	< 0.1	X	< 0.1	5.1
					EA2	X	62.0	**62.0**	X	< 0.1	< 0.1	5.3
					FAC	86.2	X	**86.2**	< 0.1	X	< 0.1	X
4	0.80	0.80	**0.72**	1.13	EA3	58.6	58.2	70.6	58.6	58.2	70.6	**90.7**
					PA2	65.8	X	65.8	65.8	X	65.8	**92.0**
					EA2	X	62.0	62.0	X	62.0	62.0	**92.3**
					FAC	71.8	X	71.8	71.8	X	71.8	X
5	**0.80**	0.82	0.80	1.22	EA3	35.2	58.2	**62.2**	22.2	47.5	50.4	58.2
					PA2	42.4	X	**42.4**	28.2	X	28.2	61.3
					EA2	X	62.0	**62.0**	X	51.5	51.5	62.0
					FAC	49.2	X	**49.2**	34.2	X	34.2	X
6	0.80	**0.75**	0.80	1.33	EA3	13.8	58.2	58.7	46.5	81.5	**82.9**	58.2
					PA2	18.4	X	18.4	54.0	X	**54.0**	61.3
					EA2	X	62.0	62.0	X	84.0	**84.0**	62.0
					FAC	23.3	X	23.3	60.7	X	**60.7**	X
7	0.90	0.85	**0.72**	0.94	EA3	42.8	13.6	44.2	76.4	32.4	77.2	**90.7**
					PA2	50.3	X	50.3	81.8	X	81.8	**92.0**
					EA2	X	15.9	15.9	X	36.0	36.0	**92.3**
					FAC	57.0	X	57.0	86.0	X	86.0	X

[a] Same assumptions as for Table 1

[b] Procedure: EA3 = Equal Allocation 3; PA2 = Proportional Allocation 2; EA2 = Equal Allocation 2;
FAC = Factorial

[c] any A = overall A or simple A effects; any B = overall B or simple B effects

[d] Numbers in bold correspond to effects of special interest

[e] X = null hypothesis not tested

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)[a]	P-value[a]
Overall effect of strict BP control	0.88 (0.73-1.06)	0.193
Simple effect of strict BP control	0.76 (0.59-0.98)	0.035
Overall effect of strict glycemia control	0.81 (0.67-0.97)	0.0254[b]
Simple effect of strict glycemia control	0.69 (0.53-0.89)	0.005[c]
Simple effect of strict BP and glycemic control	0.73 (0.56-0.94)	0.016[d]

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

[b] Significant with respect to the Proportional Allocation 2 procedure.

[c] Significant with respect to the Equal Allocation 3 and Equal Allocation 2 procedures.

[d] Significant with respect to the Equal Allocation 3, Proportional Allocation 2, and Equal Allocation 2 procedures.

Table 4: Simulation Study Results

						Familywise Error Rate (%)[a]					
sim	n	#events	censoring%	scale	shape	EA3	PA2	EA2	S_A^{str}	S_A	S_{AB}
[b]1	1000	157	84.2	0.015	1.12	4.93	4.88	4.93	4.90	4.82	5.10
2	1000	167	83.3	0.05	0.5	4.82	4.96	4.67	5.05	4.77	4.69
3	500	248	50.4	0.2	0.5	5.61	5.47	5.69	5.10	5.51	5.46
4	1000	496	50.4	0.2	0.5	5.30	5.29	5.20	5.35	5.20	4.89
5	500	208	58.4	0.01	2.0	5.57	5.61	5.49	5.70	5.53	5.40
6	1000	416	58.4	0.01	2.0	5.08	5.00	5.11	5.26	5.06	5.26
7	500	394	21.2	0.5	0.5	5.22	5.11	5.51	5.11	5.17	5.58
8	1000	788	21.2	0.5	0.5	5.03	5.12	5.12	5.11	5.06	5.13
9	500	373	25.4	0.03	2.0	4.96	5.01	5.44	5.13	5.41	5.53
10	1000	747	25.3	0.03	2.0	5.42	5.50	5.39	5.39	5.17	5.17

[a] Reports the FWE rate for the Equal Allocation 3 (EA3), Proportional Allocation 2 (PA2), and Equal Allocation 2 (EA2) multiple testing procedures. Also reports the rates that $|S_A^{\text{str}}|, |S_A|, |S_{AB}|$, respectively, are greater than 1.96, where S_A^{str}, S_A, and S_{AB} are the logrank test statistics.

[b] Shape and scale parameters chosen by fitting a Weibull nuisance hazard function to the ACCORD BP trial data.