Testing Multiple Endpoints using Alpha-Investing

Dean P. Foster and Robert A. Stine*

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

October 19, 2011

Abstract

Alpha-investing is a new way to control the number of incorrectly rejected null hypotheses in a sequence of tests. This framework accommodates tests with multiple endpoints and allows various customized designs, such as different types of gatekeeping strategies. The design of the test procedure can incorporate scientific insight when available or treat the hypothesis in the anonymous fashion of step-down tests. Detailed examples illustrate alpha-investing in gatekeeping, with applications that show its ability to mimic step-down testing and to utilize dependent tests.

Key words: alpha-spending, clinical trial, false discovery rate, gatekeeping, multiple testing

{**ras:** *TODO: Add cite to the recent Yoon paper in Stat in Med.* }

*All correspondence regarding this manuscript should be directed to Prof. Stine at the address shown with the title. He can be reached via e-mail at stine@wharton.upenn.edu.

1 Introduction

Alpha-investing controls the expected number of incorrectly rejected hypotheses when testing a sequence of hypotheses. The experimenter decides the order of the tests and can adjust the ordering as outcomes are observed. Alpha-investing resembles alpha-spending (DeMets and Lan, 1994), but has a novel difference. As in alpha-spending, each test of a hypothesis consumes some of the allotted probability for Type I errors, the alpha-level. The difference occurs when a test rejects. Tests that reject the null hypothesis add to the alpha-level available for subsequent tests and thereby improve the power of the remaining tests. Thus rejections beget more rejections. Rather than treat each test as an expense that consumes its Type I error rate, alpha-investing in effect treats tests as investments, motivating our choice of name. The framework of alpha-investing is quite broad, including as special cases step-down testing and methods related to Simes (1986). Alpha-investing also accommodates both serial and parallel gatekeeping (Dmitrienko, Offen and Westfall, 2003; Dmitrienko and Tamhane, 2007). All of these variations on alpha-investing provide uniform control of the expected false discovery rate (mFDR).

Two brief examples convey the nature of alpha-investing. Later examples fill in the details. All of the designs considered here are nonsequential. Consider first a clinical trial with primary and secondary null hypotheses. Moyé (2000) describes the trial of a medication for congestive heart failure with a single primary hypothesis H_p (mortality) and three secondary hypotheses (H_{s_1} heart function, H_{s_2} heart improvement, and H_{s_3} hospitalization). Assume that the overall error rate $\alpha = 0.05$. In an alpha-spending rule, testing H_p at the 0.05 level consumes all of the available error rate and precludes testing a secondary hypothesis. That leaves several choices for those who want to test secondary hypotheses, such as: spread the alpha-level over the four hypotheses in the fashion of a Bonferroni procedure (and suffer the resulting loss of power for testing H_p), increase the overall Type I level above the customary 0.05 threshold, or adopt a different method of testing, such as a method that controls FDR. Alpha-investing offers a compromise that controls the error rate at 0.05, but offers power for testing secondary hypotheses if a test with $\alpha = 0.05$ rejects the primary hypothesis.

Alpha-investing begins with an initial allowance for Type I errors, the *alpha-wealth* of the procedure. The alpha-wealth fluctuates as testing proceeds. Testing ends when the alpha-wealth reaches 0 or all hypotheses have been rejected. Consider the clinical trial of CHF with an initial alpha-wealth equal to 0.05. Suppose that one wants to test secondary hypotheses H_{s_j} only when H_p is rejected. In this situation, the procedure would operate as follows. First, test H_p with $\alpha = 0.05$; that is, "invest" the entire alpha-wealth in the test of the primary hypothesis. If H_p is not rejected, the invested alpha-wealth has been spent and the procedure terminates. If the test rejects H_p, however, the investment in the primary test earns 0.05 toward the alpha-wealth available for testing the secondary hypotheses. One might then, for instance, test each secondary hypothesis at level 0.05/3 or continue sequentially. Other strategies for alpha-investing allow tests of the secondary hypotheses regardless of whether the procedure rejects the primary hypothesis. D'Agostino (2000) and Turk and Dworkin (2008) offer reasons for always testing secondary hypotheses whereas others, such as O'Neill (1997), disagree. To reserve some alpha-level for testing secondary hypotheses in case the test does not reject H_p, one would invest only a portion of the initial alpha-wealth in the test of H_p. For example one could test H_p at $\alpha = 0.035$, reserving 0.015 of the alpha-wealth for testing secondary hypotheses if H_p is not rejected. Rejecting

H_p would increase the alpha-wealth to $0.015 + 0.05 = 0.065$ for testing the secondary hypotheses. Section 4 continues this example with several sets of test outcomes. It is worth noting that all of these variations of the testing procedure appeal to the same theorem which shows they control the mFDR. There is no need for a new theorem to cover each special case.

The data analysis plan for the clinical study should describe the alpha-investing strategy. Diagrams such as the one shown in Figure 1 work well for this purpose and suggest the flexibility of this approach. Nodes of the tree in the graph identify the null hypothesis to be tested and give the alpha-level for each test. This graph also tracks the alpha-wealth. Figure 1 displays a strategy for alpha-investing with one primary hypothesis H_p and three secondary hypotheses, one of which is distinguished from the other two. The procedure begins by investing $\alpha = 0.035$ in the initial test. If the initial test does not reject H_p, the left branch shows that the analysis will then test each of the secondary hypotheses at level $\alpha = 0.015/3$ without re-investing the wealth earned in these tests. (One could reinvest the wealth to obtain higher power.) If the initial test rejects H_p, then the right branch tests the specific secondary hypothesis H_{s_1} at level 0.05. The outcome of the test of H_{s_1} determines the level for the remaining two tests, which are tested without re-investing.

Our second introductory example illustrates how alpha-investing incorporates scientific input in the design of a test. Consider the comparison of two vaccines based on their ability to produce, say, four antigens. Suppose further that the molecular structure of the vaccines suggests that one antigen is particularly likely to reveal a difference between the vaccines. Conventional multivariate tests (such as an F-test or T^2 test) treat the antigens symmetrically and do not directly incorporate *a priori* insight; alpha-investing does. Order the four hypotheses of no difference between the vaccines as H_1, H_2, H_3, and H_4, with H_1 specifying the antigen most expected to show a difference and H_4 the antigen least expected to show a difference. Depending on the strength of this ordering of the antigens, invest a fraction of the initial alpha-wealth in the test of H_1. For example, one might believe that if the experiment does not reject H_1, there's little chance of rejecting H_2, H_3, or H_4. In this case, one would invest most of the alpha-wealth in the test of H_1, reserving little for testing the remaining hypotheses if H_1 is not rejected. Suppose 0.04 out of the initial alpha-wealth 0.05 is used to test H_1. If H_1 is rejected, then the procedure has alpha-wealth $0.01 + 0.05 = 0.06$ for testing H_2, H_3, and H_4. If H_1 is not rejected, one has the remaining alpha-wealth 0.05-0.04=0.01 for testing the remaining hypotheses. If the external knowledge that orders the hypotheses is accurate, then alpha-investing performs like a weighted testing procedure that knows which hypotheses are false (Foster and Stine, 2008). If the external knowledge is weak, alpha-investing can mimic an FDR analysis, as illustrated in Section 4.

The remainder of this paper develops as follows. Section 2 reviews several popular methods and terminology used in testing multiple hypotheses. Section 3 introduces alpha-investing. Section 4 presents detailed examples, including an example in which the tests are dependent. Section 5 concludes with a brief discussion.

2 Criteria for Testing Multiple Hypotheses

Assume we wish to test m null hypotheses $\mathcal{H}(m) = H_1,\ H_2,\ \ldots H_m$. Each hypothesis specifies an associated parameter θ_j (which can be a scalar or vector), and for convenience set $H_j : \theta_j = 0$. Let $\theta = (\theta_1, \ldots, \theta_m)$ denote the full set of parameters for which $\theta \in \Theta$. We address the distinction among primary and secondary hypotheses in our examples in Section 4. For the moment, consider all m hypotheses equally relevant. Two sets of indicators identify the hypotheses that are rejected and whether the rejection decision is correct. Let p_j be the p-value obtained when testing H_j at level α_j, and define the observable indicator R_j as

$$R_j = \begin{cases} 1, & \text{if } H_j \text{ is rejected } (p_j \leq \alpha_j), \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let the random variable V_j^θ indicate incorrect rejections,

$$V_j^\theta = \begin{cases} 1, & \text{if } H_j \text{ is true and } R_j = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The sum $R = \sum R_j$ denotes the total number of hypotheses rejected by a procedure, and $V^\theta = \sum V_j^\theta \leq R$ denotes the unobserved number of falsely rejected hypotheses.

One criterion for multiple testing controls the chance for any false rejection. The *family-wise error rate* (FWER) is the probability of falsely rejecting any null hypothesis from $\mathcal{H}(m)$,

$$\text{FWER} \equiv \sup_{\theta \in \Theta} \mathbb{P}_\theta(V^\theta \geq 1) \ . \tag{3}$$

An important type of control of FWER obtains under the complete null hypothesis. The *complete null hypothesis* is an important special case in which all m null hypotheses in $\mathcal{H}(m)$ are true ($\theta = 0$). In this special case, $V^\theta \equiv R$ and FWER reduces to $\mathbb{P}_0(V^\theta \geq 1) \leq \alpha$, where $\mathbb{P}_0$ denotes the probability measure under the complete null hypothesis and α denotes the experimental error rate, usually 0.05. We refer to controlling FWER under the complete null hypothesis as controlling FWER in the *weak sense.* A procedure for which FWER $< \alpha$ regardless of the truth of the hypotheses controls FWER in the *strong sense.*

Bonferroni tests control FWER in the strong sense. A Bonferroni testing procedure allocates the Type I error rate over the m hypotheses, assigning the levels so that

$\sum \alpha_j = \alpha$. Most often, the levels are equal, rejecting H_j if $p_j \le \alpha/m$. Unequal allocations of the Type I error rate allow the researcher to invest more power in selected tests, presumably those of most scientific interest. This type of allocation anticipates the construction of alpha-investing rules. The method of Holm (1979) improves the power of Bonferroni tests while retaining strong control of FWER, but the gains in power are generally slight.

To obtain substantially more power in the presence of a rejected hypothesis, Benjamini and Hochberg (1995) (BH) introduced a different criterion, the false discovery rate (FDR). FDR is the expected proportion of incorrectly rejected hypotheses (false positives) among rejected hypotheses,

$$\mathrm{FDR} = \mathbb{E}\left(\frac{V^\theta}{R} \mid R > 0\right) \mathbb{P}\left(R > 0\right) . \tag{4}$$

Under the complete null hypothesis, $V^\theta \equiv R$ and FDR $= \mathbb{P}_0(R > 0) =$ FWER. Thus, testing procedures that control FDR weakly control FWER. If the complete null hypothesis is rejected, FDR controls the proportion of false positives among the rejected hypotheses. Since FDR decreases as the number of false null hypotheses increases (Dudoit, Shaffer and Boldrick, 2003), FDR becomes more easy to control in the presence of non-zero effects. This property of FDR allows more powerful testing procedures such as the step-down procedure of BH. Variations on FDR include pFDR (which drops $\mathbb{P}\left(R > 0\right)$ from (4); see Storey, 2002, 2003) and the local false discovery rate fdr(z) (which estimates the false discovery rate; see Efron, 2007, 2010). Closer to alpha-investing, Meinshausen and Bühlmann (2004) and Meinshausen and Rice (2006) estimate the number of false null hypotheses in $\mathcal{H}(m)$, and Genovese, Roeder and Wasserman (2006) weight p-values based on prior knowledge that identifies hypotheses that are likely to be false.

Alpha-investing controls a similar error rate. Rather than control the expected value of the ratio V^θ/R, alpha-investing controls the ratio of the expected values known as mFDR. A testing procedure controls mFDR at level α if

$$\text{mFDR} \equiv \frac{\mathbb{E}\, V^\theta}{\mathbb{E}\, R + 1} \leq \alpha. \tag{5}$$

mFDR traditionally does not typically include the 1 in the denominator. The addition of a positive constant to the denominator is necessary under the complete null hypothesis; under the complete null hypothesis $V^\theta \equiv R$. Control of mFDR implies weak control of FWER. Under the complete null hypothesis, mFDR $\leq \alpha$ implies that FWER $\leq \frac{\alpha}{1-\alpha}$. Benjamini and Hochberg (1995) considered mFDR, but viewed this criterion as artificial because it controls a ratio of expectations rather than a property of the realized sequence of tests. In practice, we find small differences in the form of control (Foster and Stine, 2008).

3 Alpha-Investing

Alpha-investing has a single tuning parameter α which determines the initial alpha-wealth as well as the gain in alpha-wealth when a test rejects. Let $W(j) \geq 0$ denote the accumulated alpha-wealth after testing j hypotheses; $W(0)$ is the initial alpha-wealth. Conventionally, $W(0) = \alpha = 0.05$. Given α, an alpha-investing rule is a function $\mathcal{A}_\alpha$ that sets the level of the next test, potentially using the observed outcomes of the $j-1$ prior tests:

$$\alpha_j = \mathcal{A}_\alpha(R_1,\, R_2, \ldots, R_{j-1}) \tag{6}$$

The only condition is that $0 \leq \alpha_j \leq W(j-1)$. Figure 1 in the introduction displays an alpha-investing rule as a directed graph. The alpha-wealth $W(j)$ fluctuations up and down as testing proceeds. Each test at level α_j reduces the available alpha-wealth by α_j, as in alpha-spending. If $p_j \le \alpha_j$, the test rejects H_j ($R_j = 1$) and the alpha-wealth increases by α. The change in the alpha-wealth at test j is then

$$W_j = W(j) - W(j-1) = \alpha\, R_j - \alpha_j\ . \tag{7}$$

For some intuition, consider independent tests under the complete null hypothesis. In this case, each p-value is uniformly distributed on [0,1], and the expected change in the alpha-wealth is $\mathbb{E}\, W_j = -\alpha_j(1-\alpha) < 0$. This suggests that the alpha-wealth steadily decreases when testing a sequence of true null hypotheses. Other investing and payment systems that offer greater flexibility are possible, albeit at the cost of adding further tuning parameters (Foster and Stine, 2008).

The order in which hypotheses are tested is entirely up to the statistician and may depend on the outcome of prior tests. The underlying theory only requires that the test of H_j, given the history of prior rejections, have level not exceeding α_j:

$$\forall \theta \in \Theta, \quad \mathbb{E}\,(V_j^\theta \mid R_1,\, R_2,\, \ldots, R_{j-1}) \le \alpha_j\ . \tag{8}$$

Note that the test of H_j is conditioned only on prior accept/reject decisions, not on test statistics (such as z-scores) or parameter estimates. Results established in Lehmacher and Wassmer (1999) and Tsiatis and Mehta (2003) suggest that one obtains a more powerful procedure by conditioning on acceptance rather than the test statistic.

An appealing feature of alpha-investing is that, in spite of its flexibility, it guarantees control of mFDR at level α. So long as the tests are “honest” in the sense of (8), alpha-investing bounds the expected number of false rejections by α times 1 plus the total number of rejections:

Theorem 1 *An alpha-investing rule $\mathcal{A}_\alpha$ that meets (6) with initial alpha-wealth and pay-out α controls mFDR at level α if the sequence of tests satisfy (8).*

Because the proof of this result relies only on the optional stopping theorem for martingales, we do not require independent tests. (For completeness, a short proof of this theorem is in the appendix.) An example in Section 4 illustrates alpha-investing with dependent tests; that example shows that dependence complicates the choice of rejection regions and affects the hypotheses that can be tested.

Alpha-investing also provides a novel type of uniform control that allows the investigator to stop the testing process before it completes. One can stop the testing at any intermediate rejection and still guarantee control of mFDR. For instance, rather than testing all of a large set of hypotheses, one is free to stop testing hypotheses after rejecting, say, three of them. Alpha-investing bounds the expected number of false rejections among these three. In particular, we have the following theorem (Foster and Stine, 2008)

Theorem 2 *Let T_r denote the position of the hypothesis at which the jth rejection occurs, and let $V^\theta(j)$ denote the number of false rejections among tests of the first j hypotheses. An alpha-investing rule $\mathcal{A}_\alpha$ has the property that $\mathbb{E}\, V^\theta(T_r) \leq \alpha(r+1)$*

4 Examples

This section gives two detailed examples of alpha-investing in the context of clinical trials with multiple endpoints. Both examples concern testing a collection of four hypotheses, of which one or two are primary hypotheses. The first clinical example (Section 4.1) continues the gatekeeping example from the introduction; this example illustrates how alpha-investing can mimic step-down testing. The second clinical example (Section 4.3) illustrates the use of alpha-investing when the tests are dependent; a simpler example in Section 4.2 introduces issues in dependent tests.

4.1 Independent tests and step-down testing

The first example tests a single primary hypothesis followed by three secondary hypotheses. To make the calculations explicit, Table 1 shows p-values from three scenarios labeled A, B, and C by Chen, Luo and Capizzi (2005). Assume that these tests are independent.

Suppose first that we intend to test the secondary hypotheses only if the primary hypothesis is rejected. In this case, alpha-investing commits the entire initial alpha-wealth $W(0) = 0.05$ to the test of the primary hypothesis. Because the p-value of the test of H_p is less than 0.05 in all three scenarios in Table 1, the procedure rejects H_p and proceeds to test the secondary hypotheses with alpha-wealth 0.05. Assume that we have no *a priori* reason to suspect any secondary hypothesis to be more likely to be rejected than another. For this situation, we use alpha-investing to implement a version of step-down testing that is achieved by revisiting prior hypotheses. First test each of H_{s_1}, H_{s_2}, and H_{s_3} at the Bonferroni level $0.05/3 \approx 0.0167$. In Scenarios A and B, this first pass rejects both H_{s_1} and H_{s_3}. Hence, the alpha-wealth available to test H_{s_2} is 0.10. (The test spent 0.05 for the three tests at the Bonferroni level, but earned 0.05 for each rejection.) With only one hypothesis remaining, the procedure can invest all of the remaining alpha-wealth in the test of H_{s_2} and reject this null hypothesis, even though the p-value in Scenario B is 0.06.

Now consider scenario C. The initial tests at level 0.0167 reject only H_{s_3}, leaving alpha-wealth 0.05 for testing H_{s_1} and H_{s_2}. Testing these at level $0.05/2 = 0.025$ appears to fail to reject either, but this is incorrect. Since we did not reject either secondary hypothesis in the first pass, both p_{s_1} and $p_{s_2} > 0.0167$. Hence, conditional on not rejecting in the first pass, the threshold for rejection in the second pass is p_1^*

which is chosen so that

$$\mathbb{P}\,(p_j \le p_1^* | p_j > 0.0167) = \frac{p_1^* - 0.0167}{1 - 0.0167} = 0.025\ , \tag{9}$$

implying $p_1^* = 0.04209$. Since $p_{s_1} = 0.03$ in scenario C, the second visit rejects H_{s_1}, earning 0.05 toward the test of H_{s_2}. For the final test, the p-value threshold increases to p_2^* that satisfies

$$\mathbb{P}_0(p_j \le p_2^* | p_j > 0.04209) = \frac{p_2^* - 0.04209}{1 - 0.04209} = 0.05\ , \tag{10}$$

or $p_2^* = 0.0943$. Since $p_{s_2} \le p_2^*$, the procedure rejects H_{s_2}. Once again, alpha-investing rejects all three secondary hypotheses. This rejection of the secondary hypotheses differs from the performance of a step-down test of these three hypotheses at level 0.05. The ordered p-values of the secondary hypotheses are 0.002, 0.03, and 0.06. The p-values of H_{s_3} and H_{s_1} are less than the corresponding step-down thresholds 0.05/3, 0.10/3, and 0.15/3. The step-down test would not reject H_{s_2} since $p_{s_2} > 0.05$.

Alternatively, one might prefer to test the secondary hypotheses even if the primary hypothesis is not rejected. This preference requires one to reserve some alpha-wealth for the secondary tests. For instance, assume that alpha-investing begins by investing $\alpha_1 = 0.035$ in the test of H_p. Because the p-value of the primary hypothesis is 0.048, the test of the primary hypothesis no longer rejects H_p, leaving $W(1) = 0.015$ for testing the secondary hypotheses. Though starting with less alpha-wealth, the alpha-investing version of step-down testing again rejects all three secondary hypotheses. In scenarios A and B, both p_{s_1} and p_{s_3} are less than the initial Bonferroni threshold $0.015/3 = 0.005$. As before, these rejections generate alpha-wealth $2(0.05) = 0.10$ for testing H_{s_2}. For scenario C, alpha-investing rejects H_{s_3} in the first pass. For the second pass, the conditional p-value threshold p_1^* is smaller than in (9) because of the smaller initial level,

$$\mathbb{P}_0(p_j \leq p_1^* | p_j > 0.005) = \frac{p_1^* - 0.005}{1 - 0.005} = 0.025 \Rightarrow p_1^* = 0.0301 \ .$$

Hence, the second pass at level 0.025 rejects H_{s_1} (just barely), leaving alpha-wealth 0.05 for the final test of H_{s_2} conditional on its p-value being larger than 0.0301. The conditional threshold is

$$\mathbb{P}_0(p_j \leq p_2^* | p_j > 0.0301) = \frac{p_2^* - 0.0301}{1 - 0.0301} = 0.05 \Rightarrow p_2^* = 0.0816 \ .$$

As before, this threshold implies rejecting H_{s_2}. One could now revisit the test of the primary hypothesis.

4.2 Dependent tests

This section introduces the use of alpha-investing with dependent tests in the simple context of testing a pair of hypotheses. The key condition is that the tests must account for dependence in the sense of (8). Dependence complicates both the construction of rejection regions as well as the formulation of the hypotheses. The complexities are not unique to alpha-investing and concern any procedure that analyzes a sequence of dependent tests.

Consider testing a pair of one-sided hypotheses $H_1 : \mu_X \leq 0$ and $H_2 : \mu_Y \leq 0$ with dependent means $\overline{X}$ and $\overline{Y}$. One-sided tests simplify the construction of rejection regions; a two-sided test can be represented as a sequence of two one-sided tests. Assume that the means are standardized so that $\overline{X}$ and $\overline{Y}$, the test statistics, have a bivariate normal distribution with means μ_X and μ_Y, equal variance 1, and correlation ρ. We assume that ρ and the variances are known.

Suppose that we test H_1 and H_2 at level $\alpha_1 = \alpha_2 = 0.05$. The threshold for testing H_1 is $\tau_1 = 1.645$. The threshold for testing H_2 depends on whether we reject H_1, so we denote this threshold as τ_{2,R_1}. For instance, $\tau_{2,0}$ is the threshold for testing H_2 given that we did not reject H_1 ($R_1 = 0$). The threshold τ_{2,R_1} must satisfy

$$\forall \mu_X \qquad \sup_{\mu_Y \le 0} \mathbb{P}\left(\overline{Y} > \tau_{2,r} | R_1 = r\right) \le \alpha_2 \ . \tag{11}$$

The difficulty in finding τ_{2,R_1} arises because accepting or rejecting H_1 does not constrain μ_X. The condition (11) must hold *for all* choices of μ_X, not just those consistent with the outcome of the test of H_1. To help appreciate the context, Figure 2 shows a contour plot of the joint distribution of $\overline{X}$ and $\overline{Y}$ with correlation $\rho = 0.5$ and means $\mu_X = \mu_Y = 0$. The diagonal line identifies the conditional mean of $\overline{Y}$ given $\overline{X}$, and the dashed line identifies the rejection region for testing H_1. If the pair $(\overline{X}, \overline{Y})$ lies to the left of the dashed line in the figure, the test does not reject H_1. In that case, to find $\tau_{2,0}$, we have to find values for μ_X and $\mu_Y \le 0$ which together maximize the conditional probability in (11). To do that, imagine sliding the joint distribution in Figure 2 horizontally to the left or right. Sliding the distribution to the left shifts the conditional means higher, placing more probability in the rejection region. If $\mu_X \ll 0$, conditioning on $R_1 = 0$ is uninformative and we see that as $\mu_X \to -\infty$, $\tau_{2,0} \to 1.645$. On the other hand, suppose $R_1 = 1$. The relevant portion of the joint distribution lies to the right of τ_1. As μ_X decreases, less and less of the joint distribution lies to the right of τ_1. Within this diminishing portion of the joint distribution, however, the conditional mean of $\overline{Y}$ given $\overline{X} > \tau_1$ increases as μ_X decreases. As a result, no finite $\tau_{2,1}$ satisfies the condition (11).

Our solution to this problem is to modify the hypotheses. To identify τ_2, we require that H_2 bound the possible values for μ_X. If the first test does not reject H_1, the second hypothesis becomes $H_2' : \mu_Y \le 0; \mu_X \le 0$. If the first test rejects H_1, then the second hypothesis becomes $H_2' : \mu_Y \le 0; \mu_X \ge 0$. With this revision to H_2, $\tau_{2,0}$ is unchanged, and $\tau_{2,1}$ satisfies

$$\sup_{\mu_X \geq 0, \mu_Y \leq 0} \mathbb{P}\,(\overline{Y} > \tau_{2,1} | R_1 = 1) = \sup_{\mu_X \geq 0, \mu_Y \leq 0} \mathbb{P}\,(\overline{Y} > \tau_{2,1} | \overline{X} > 1.645) \leq \alpha_2 \ . \tag{12}$$

The supremum occurs on the boundary with $\mu_X = 0$, and numerical integration gives $\tau_{2,1} = 2.495$.

The sequence of hypotheses must accumulate in this fashion in the presence of dependent tests in order to guarantee a viable rejection region. It is likely that many users are inclined to interpret the results of a sequence of hypothesis tests in this conditional fashion, irrespective of how the hypotheses are stated. Alpha-investing requires this formality in how we state hypotheses.

4.3 Dependent testing in practice

To illustrate these calculations in a gatekeeping application, consider testing four hypotheses using dependent tests. Zhang, Quan and Ng (1997) describe a clinical trial that compares a drug for asthma to placebo. Table 2 summarizes a similar, but imaginary, clinical trial that compares mean responses of 35 treated subjects to 35 subjects who received placebo. (We modified the effect sizes from those in Zhang et al. (1997) for this illustration so that a test fails to reject, but retained the correlations.) Larger means indicate greater efficacy. Table 2 includes the pooled two-sample standard error of the mean treatment versus placebo (SE) and the correlations between outcomes. For example, the correlation between the measured symptom score and medication use is 0.67. To simplify the calculations, we use a normal model for the sampling distribution of the means; otherwise, one would integrate a multivariate t-distribution. The use of a normal distribution rather than a t-distribution would change the p-values reported in Zhang et al. (1997) in the third decimal place.

Suppose that the analysis is designed to proceed as shown in Figure 3. The hypotheses are one-sided of the form $H_j : \mu_j \leq 0$, numbered as in Table 2. Rejecting H_j implies finding a statistically significant clinical effect. The initial wealth is $W(0) = 0.05$ and each rejection earns 0.05 toward testing subsequent hypotheses. The design treats H_1 and H_2 as primary hypotheses and H_3 and H_4 as secondary. We first test whether either expiratory volume or flow rate reject before testing the symptom score or medication use, and we allocate $\alpha_1 = 0.025$ and $\alpha_2 = 0.025$. If neither test rejects, then the alpha-wealth falls to zero and we cannot test H_3 or H_4 (the far left branch in Figure 3). So long as either primary hypothesis is rejected, the design guarantees that the alpha-wealth available for testing H_3 and H_4 is $W(2) = 0.05$ or 0.10.

For the first test, the familiar one-sided z-test rejects H_1 since $z_1 > \tau_1 = 1.96$ (see Table 2), and the alpha-wealth increases to $W(1) = 0.075$. Because we reject H_1, we could test H_2 using an alpha-level greater than $\alpha_2 = 0.025$ and still guarantee some alpha-wealth would remain for testing secondary hypotheses. We assume here, however, that the study follows the plan in Figure 3 and fixes the levels for testing H_1 and H_2 at $\alpha_1 = \alpha_2 = 0.025$.

The remaining tests must account for the dependence. As in the example in Section 4.2, since we reject H_1, the second hypothesis becomes $H_2' : \mu_2 \leq 0;\ \mu_1 \geq 0$. The threshold $\tau_{2,1}$ must satisfy

$$\sup_{\mu_1 \geq 0,\, \mu_2 \leq 0} \mathbb{P}\left(Z_2 > \tau_{2,1} \mid Z_1 > \tau_1\right) = 0.025\ .$$

Solving numerically, we obtain $\tau_{2,1} = 2.49$ (which is coincidently similar to the choice of $\tau_{2,1}$ in the example of Section 4.2). The small correlation 0.25 between z_1 and z_2 produces a substantial impact on the threshold for testing H_2; were the tests independent, then $\tau_2 = 1.96$. Since $z_2 = 1.82 < \tau_{2,1}$, we do not reject H_2'. (One could revisit H_2 by investing some of the remaining alpha-wealth in testing this hypothesis as described in Section 4.1, but we continue to the remaining hypotheses.)

Since we reject H_1 but not H_2, the alpha-wealth remaining after testing the primary hypotheses is $W(2) = 0.05$. To reserve some alpha-wealth for testing H_4, we test H_3 using half of the available wealth. The accumulated third hypothesis is $H_3' : \mu_3 \leq 0;\ \mu_1 \geq 0, \mu_2 \leq 0$ with $\alpha_3 = 0.025$. The threshold $\tau_{3,10} = 2.601$ satisfies the condition that

$$\sup_{\mu_1 \geq 0,\, \mu_2 \leq 0,\, \mu_3 \leq 0} \mathbb{P}\left(Z_3 > \tau_{3,10} \mid Z_1 > \tau_1, Z_2 \leq \tau_{2,1}\right) = 0.025 \ .$$

Since $z_3 > \tau_{3,10}$ $(3.13 > 2.601)$, we reject H_3'. The alpha-wealth grows to $W(3) = 0.075$, which we can spend on the final test of $H_4' : \mu_4 \leq 0; \mu_1 \geq 0, \mu_2 \leq 0, \mu_3 \geq 0$. The final threshold $\tau_{4,101} = 1.964$ satisfies

$$\sup_{\mu_1 \geq 0,\, \mu_2 \leq 0,\, \mu_3 \geq 0} \mathbb{P}\left(Z_4 > \tau_{4,101} \mid Z_1 > \tau_1, Z_2 \leq \tau_{2,1}, Z_3 > \tau_{3,10}\right) = 0.075 \ .$$

Hence, we do not reject H_4' since $z_4 < \tau_{4,101}$ $(1.75 < 1.964)$.

5 Discussion

Research has produced a variety of tests for multiple endpoints. Whether performing an interim test (as in Kieser, Bauer and Lehmacher, 1999) or implementing gatekeeping (e.g. Dmitrienko and Tamhane, 2007), the objective of many of these has been to provide strong control of the FWER. This is usually achieved by developing Bonferroni-Holm bounds that distribute the fixed alpha-level over the hypotheses in a way that produces a closed testing procedure in the sense of Marcus, Peritz and Gabriel (1976). Such closed designs can be difficult to obtain and describe, as evident in the figures of Kieser et al. (1999) and the graphical methodology of Bretz, Maurer, Brannath and Posch (2009). Alpha-investing offers an arguably simpler approach, at the cost of trading strong control of the FWER for weak control of the FWER and control of the expected false discovery rate. The use of methods that control FDR or mFDR may be inappropriate in some cases, such as those intended for regulatory approval, as noted by Neuhäuser (2006). We suspect that, as in genetics, alternative error rates such as FDR and mFDR will become increasingly accepted as the prevalence of studies with numerous hypotheses increases in clinical studies.

We plan to explore the use of alpha-investing for monitoring sequential trials that often have multiple endpoints. Alpha-spending rules have been used to control the error rates in these trials, including designs with primary endpoints (for example, Kosorok, Shi and DeMets, 2004) and, more recently, designs with both primary and secondary endpoints (Tamhane, Mehta and Liu, 2010). The optional stopping properties of alpha-investing highlighted in Theorem 2 suggest that we can establish the same guaranteed control of mFDR in the sequential context as obtained here for non-sequential trials. The ability of alpha-investing to revisit a hypothesis should allow us to view the sequential tests as a martingale. This characterization might simplify the analysis of experiments that stop collecting data based on one endpoint, but still want to test secondary endpoints.

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statist. Soc., Ser. B*, **57**, 289–300.

Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009) A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, **28**, 586–604.

Chen, X., Luo, X. and Capizzi, T. (2005) The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine*, **24**, 1385–1397.

D'Agostino, R. B. (2000) Controlling alpha in a clinical trial: the case for secondary endpoints. *Statistics in Medicine*, **19**, 763–766.

DeMets, D. L. and Lan, K. G. (1994) Interim analysis: The alpha spending function approach (Disc: p1353-1356). *Statistics in Medicine*, **13**, 1341–1352.

Dmitrienko, A., Offen, W. W. and Westfall, P. H. (2003) Gatekekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*, **22**, 2387–2400.

Dmitrienko, A. and Tamhane, A. C. (2007) Gatekekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, **6**, 171–180.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Efron, B. (2007) Size, power, and false discovery rates. *Annals of Statistics*, **35**, 1351–1377.

— (2010) *Large Scale Inference: Empirical Bayes methods for estimation, testing, and prediction*. New York: Cambridge.

Foster, D. P. and Stine, R. A. (2008) α-investing: a procedure for sequential control of expected false discoveries. *JRSS-B*, **70**, 429–444.

Genovese, C., Roeder, K. and Wasserman, L. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Kieser, M., Bauer, P. and Lehmacher, W. (1999) Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*, **41**, 261–277.

Kosorok, M. R., Shi, Y. and DeMets, D. L. (2004) Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics*, **60**, 134–145.

Lehmacher, W. and Wassmer, G. (1999) Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286–90.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Meinshausen, N. and Bühlmann, P. (2004) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence. *Biometrika*, **92**, 893–907.

Meinshausen, N. and Rice, J. (2006) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, **34**, 373–393.

Moyé, L. A. (2000) Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine*, **19**, 767–779.

Neuhäuser, M. (2006) How to deail with multiple endpoints in clinical trials. *Fundamental & Clinical Pharmacology*, **20**, 515–523.

O'Neill, R. T. (1997) Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*, **18**, 550–556.

Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statist. Soc., Ser. B*, **64**, 479–498.

— (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

Tamhane, A. C., Mehta, C. R. and Liu, L. (2010) Testing a primary and a secondary endpoint in a group sequential design. *Biometrics*, **66**, 1174–1184.

Tsiatis, A. A. and Mehta, C. (2003) On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**, 367–378.

Turk, D. C. and Dworkin, R. H. *et al.* (2008) Analyzing multiple endpoints in clinical trials of pain treatments: IMPACT recommendations. *Pain*, **139**, 485–493.

Zhang, J., Quan, H. and Ng, J. (1997) Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials*, **18**, 204–221.

Appendix

The definition of alpha-investing in Section 3 differs slightly from that in Foster and Stine (2008). To show that the same properties obtain, we verify that a specific sequence of random variables forms a submartingale. This sequence combines the observed random variables R_j from (1) and V_j^θ from (2). The V_j^θ are not observable as they depend on the unknown parameters θ. The cumulative sums of these indicators are $R(j) = \sum_{i=1}^j R_j$ and $V^\theta(j) = \sum_{i=1}^j V_j^\theta$. Similarly, $W(j)$ denotes the accumulated alpha-wealth after j tests. The specific sequence of random variables that define the performance of alpha-investing is

$$A^\theta(j) = \alpha(R(j)+1) - V^\theta(j) - W(j) \ , \quad A^\theta(0) = 0. \tag{13}$$

If $A^\theta(j)$ is a submartingale, then

$$\mathbb{E}\,\left(A^\theta(j) \mid A^\theta(j-1), A^\theta(j-2), \ldots, A^\theta(1)\right) \geq \mathbb{E}\,A^\theta(0) = 0\ . \tag{14}$$

Hence, $\mathbb{E}\,V^\theta(j) + W(j) \leq \alpha(\mathbb{E}\,R(j)+1)$, and the alpha-investing rule controls mFDR at level α.

To show that $A^\theta(j)$ is a submartingale, consider the conditional expectation of

$$\begin{aligned} A_j^\theta &= A^\theta(j) - A^\theta(j-1) \\ &= \alpha R_j - V_j^\theta - W_j \\ &= \alpha_j - V_j^\theta \ , \end{aligned} \tag{15}$$

where the last step uses (7). If H_j is false, then $V_j^\theta = 0$ so that $A_j^\theta = \alpha_j \geq 0$. If H_j is true, then

$$\mathbb{E}\,\left(V_j^\theta \mid A^\theta(j-1), A^\theta(j-2), \ldots, A^\theta(1)\right) \leq \alpha_j \tag{16}$$

since the level of the test of H_j is α_j. Because the sigma field generated by $R_1, \ldots, R_m$ is equivalent to the sigma field generated by $A^\theta(1), \ldots, A^\theta(m)$ (the parameters θ are fixed), the condition that each test has level α_j in (8) implies (16).

Table 1: *P-values for a primary hypothesis and three secondary hypotheses* $\alpha = 0.05$ *(from Chen et al., 2005).*

Hypothesis		p-value scenario A	 B	 C
Primary	H_p	0.048	0.048	0.048
Secondary	H_{s_1}	0.003	0.003	0.030
	H_{s_2}	0.026	0.060	0.060
	H_{s_3}	0.002	0.002	0.002

Table 2: *Summary statistics of asthma trial. Adapted from a pooled two-sample comparison of treatment to placebo for 4 outcomes (Zhang et al., 1997).*

Endpoint	z	Mean	SE	Correlations			
1. Expiratory volume	2.36	6.50	2.75	1.00	0.25	0.31	0.24
2. Expiratory flow rate	1.82	9.76	5.35	0.25	1.00	0.42	0.43
3. Symptom score	3.13	0.72	0.23	0.31	0.42	1.00	0.67
4. Medication use	1.75	0.28	0.16	0.24	0.43	0.67	1.00

Figure 1: *Alpha-investing requires a stated strategy for how the testing will proceed, allocating accumulated alpha-wealth $W(j)$ over the remaining hypotheses. This example shows a possible strategy for testing a primary hypothesis H_p and three secondary hypotheses. Left branches indicate "accepts" that reduce the alpha-wealth; right branches are "rejects" that increase the alpha-wealth.*

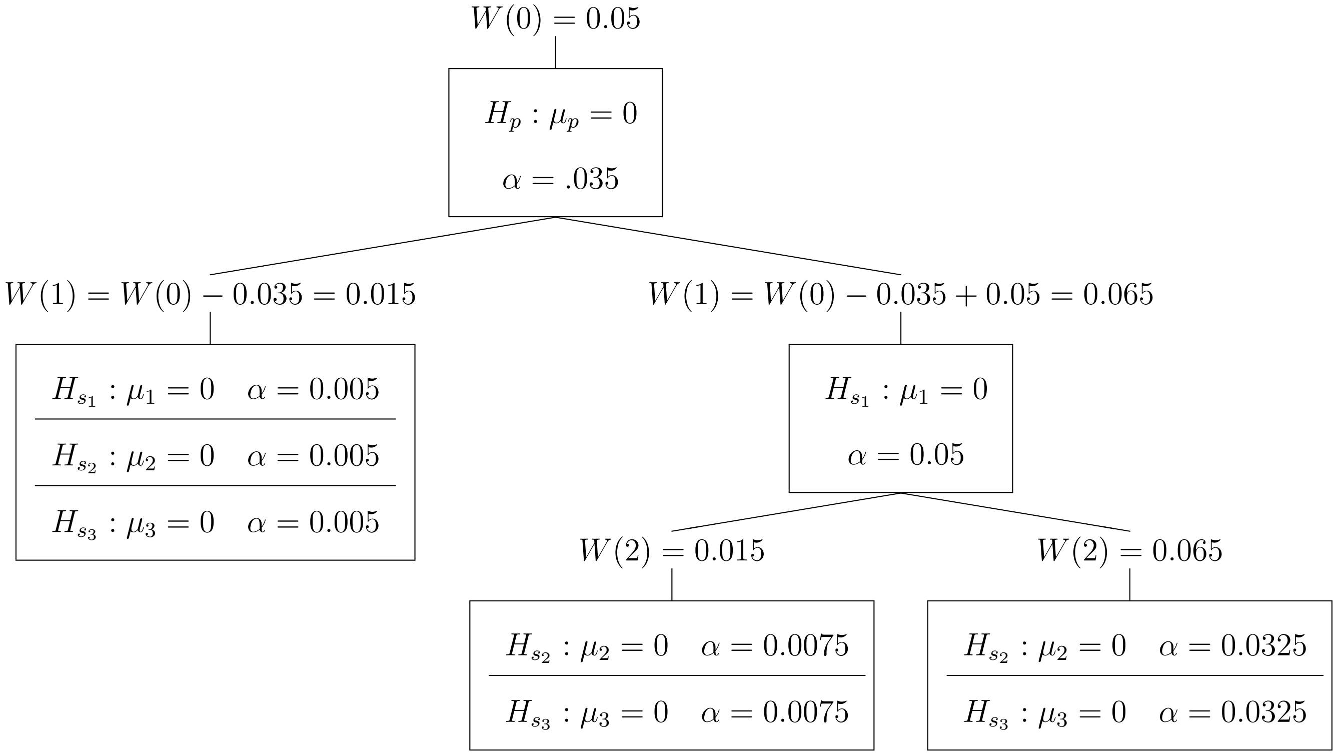

$W(0) = 0.05$
$H_p : \mu_p = 0$
$\alpha = .035$
$W(1) = W(0) - 0.035 = 0.015$
$W(1) = W(0) - 0.035 + 0.05 = 0.065$
$H_{s_1} : \mu_1 = 0 \quad \alpha = 0.005$
$H_{s_2} : \mu_2 = 0 \quad \alpha = 0.005$
$H_{s_3} : \mu_3 = 0 \quad \alpha = 0.005$
$H_{s_1} : \mu_1 = 0$
$\alpha = 0.05$
$W(2) = 0.015$
$W(2) = 0.065$
$H_{s_2} : \mu_2 = 0 \quad \alpha = 0.0075$
$H_{s_3} : \mu_3 = 0 \quad \alpha = 0.0075$
$H_{s_2} : \mu_2 = 0 \quad \alpha = 0.0325$
$H_{s_3} : \mu_3 = 0 \quad \alpha = 0.0325$

Figure 2: *Finding the rejection region for a dependent test of H_2 at level $\alpha_2 = 0.05$ requires locating a threshold so that at most 5% of the distribution to the left or right of τ_1 lies above $\tau_{2,0}$ or $\tau_{2,1}$, conditional on whether $\overline{X}$ is greater or less than τ_1.*

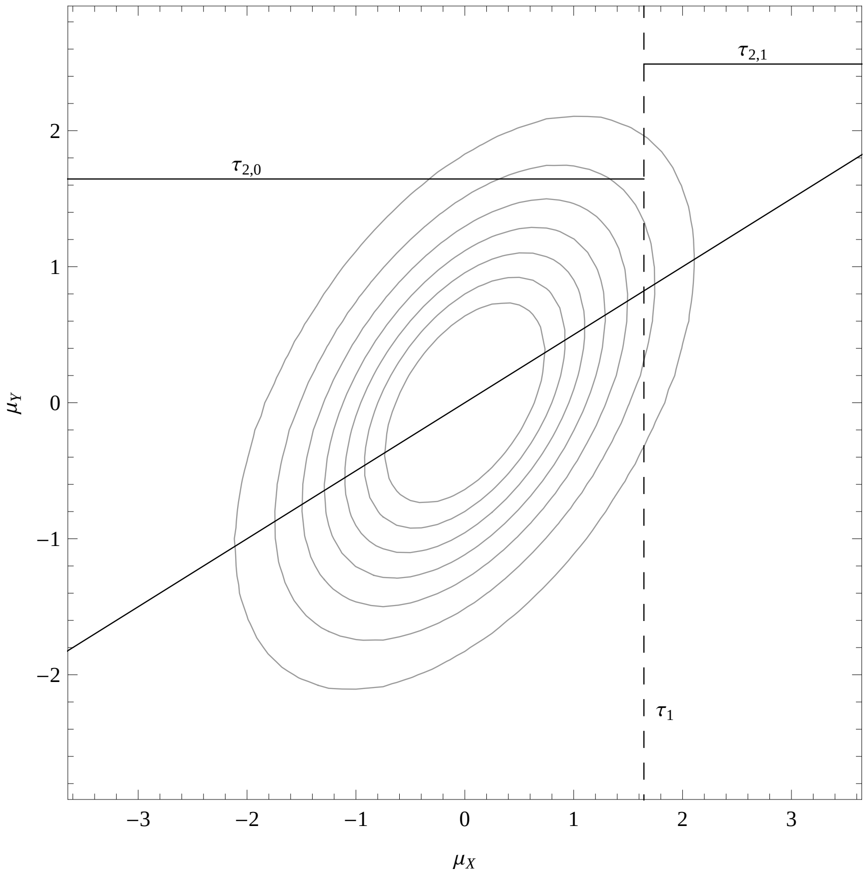
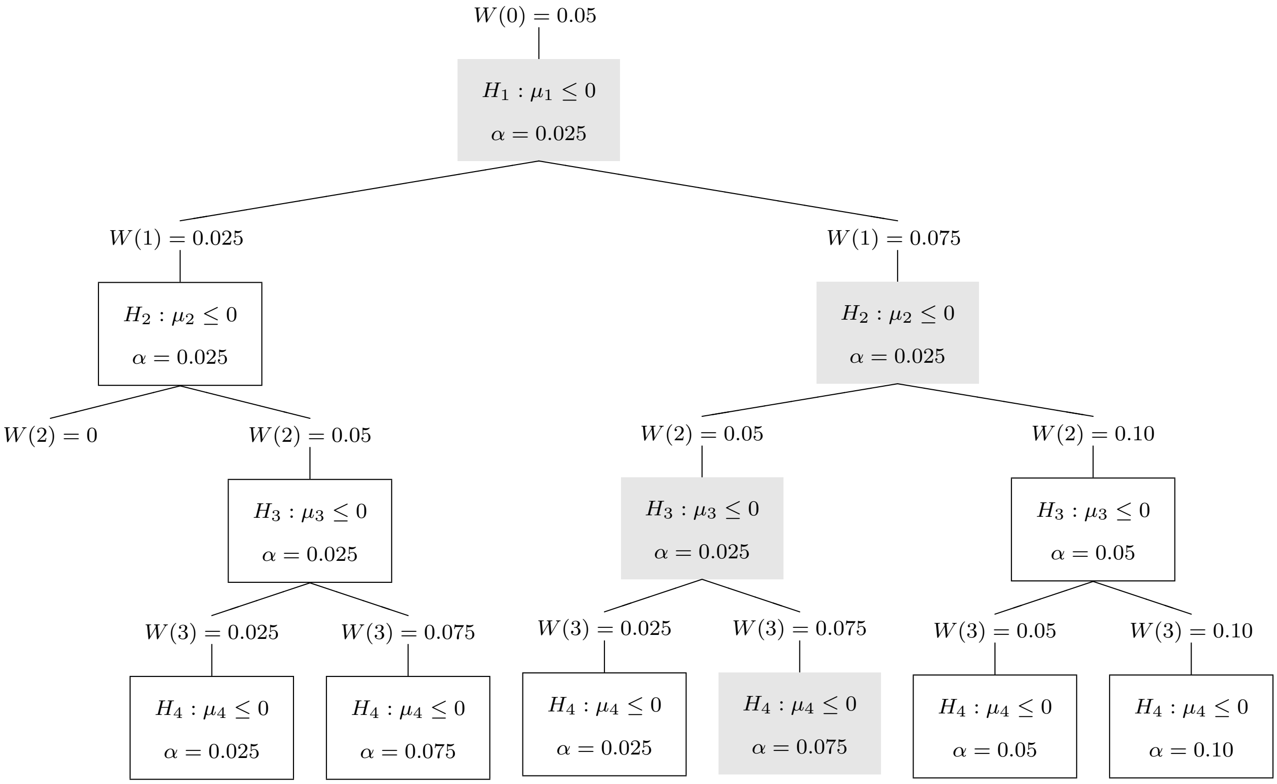


Figure 3: *Alpha-investing design for testing two primary hypotheses H_1 and H_2 and two secondary hypotheses H_3 and H_4. Shaded nodes indicate the path followed in the example; the data reject H_1 and H_3.*


$W(0) = 0.05$
$H_1 : \mu_1 \leq 0$
$\alpha = 0.025$
$W(1) = 0.025$
$W(1) = 0.075$
$H_2 : \mu_2 \leq 0$
$\alpha = 0.025$
$H_2 : \mu_2 \leq 0$
$\alpha = 0.025$
$W(2) = 0$
$W(2) = 0.05$
$W(2) = 0.05$
$W(2) = 0.10$
$H_3 : \mu_3 \leq 0$
$\alpha = 0.025$
$H_3 : \mu_3 \leq 0$
$\alpha = 0.025$
$H_3 : \mu_3 \leq 0$
$\alpha = 0.05$
$W(3) = 0.025$
$W(3) = 0.075$
$W(3) = 0.025$
$W(3) = 0.075$
$W(3) = 0.05$
$W(3) = 0.10$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.025$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.075$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.025$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.075$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.05$
$H_4 : \mu_4 \leq 0$
$\alpha = 0.10$