and they were combined into one for logistic/efficiency reasons. Moreover, the t-statistics for the effects of antioxidants and HRT are nearly statistically independent. If we knew the population variances and substituted them for their estimates in the denominators of the t-statistics, the test statistics would be completely independent. The analogy with separate clinical trials is strong. Consequently, no multiple comparison adjustment was made for the two comparisons.

Table 4.6: The Women's Angiographic Vitamin and Estrogen (WAVE) trial was a 2×2 factorial trial of antioxidants and hormone replacement therapy (HRT) in postmenopausal women with heart disease. Women were randomized in equal proportions to the four cells.

	Placebo HRT	Active HRT
Placebo antioxidant		
Active antioxidant		

Factorial trials can have more than two levels of each factor or more than two factors. For example, if the WAVE trial had included 3 levels of antioxidants (placebo, dose 1, and dose 2) and two levels of HRT, it would have been a 3×2 factorial. If it had instead combined a trial of another factor such as diet A versus diet B, it would have been a $2 \times 2 \times 2$ factorial trial.

A different multi-arm design compares k treatments to the same control arm. An example is the Antihypertensive and Lipid-Lowering treatment to reduce Heart Attack (ALLHAT) trial described in Example 4.6. Three newer classes of antihypertensive drugs were compared to a diuretic with a longer track record. Here, the $k = 3$ comparisons with the diuretic are not statistically independent because the same control arm is used for all comparisons. A randomly 'bad' control arm might increase the probability of a false positive for each comparison. For this reason, clinical trialists have traditionally made a multiple comparison adjustment in this type of trial, as opposed to in a factorial trial. That thinking began to change in the era of emerging infectious diseases like Ebola virus disease and COVID-19. One reason is that in a deadly pandemic, there is an urgent need to find something that works. Adhering to an arguably unnecessary, austere level of rigor is counterproductive. Moreover, pharmaceutical companies would have no incentive to join a multi-armed trial that adjusts for multiple comparisons when they can perform their own trial with no multiplicity adjustment. Still, if more than one dose of the same drug is compared to a control, a multiple comparison adjustment is in order if a claim of efficacy could be made based on at least one dose being beneficial. The simplest and most conservative multiple comparison adjustment is the *Bonferroni method* with significance level α/k for each of the k comparisons. This is based on the *Bonferroni inequality* $P\{\cup_i A_i\} \leq \sum_i P(A_i)$ for any events $A_1, A_2, \ldots$ Chapter 10 contains more powerful alternative methods.

Another thing to bear in mind for comparisons with the same control arm is that the sample size in the control arm is often made larger than the sample size in each other arm. A heuristic explanation is as follows. If a patient is assigned to one of the other arms, power for the comparison of that arm with control is increased, but if the patient is assigned to the control arm, power for each comparison with control is increased. Of course, there is a limit to how large the control arm should be; if all patients are assigned to control, there are no comparisons! Likewise, if almost all patients are assigned to control, power is poor.

We can determine the allocation ratio that maximizes power for each comparison with control as follows. Suppose that N patients are randomized to k comparator arms plus a

control (a total of $k+1$ arms) in a trial comparing means. Let n_0 and n denote the sample sizes in the control and each of the other arms, respectively. The constraint is that

$$n_0 + kn = N. \tag{4.17}$$

Assume a common known variance σ^2 in the different arms. Power for the comparison of arm i to control (arm 0) is a decreasing function of $\text{var}(\bar{Y}_i - \bar{Y}_0) = \sigma^2(1/n + 1/n_0)$. We would like to choose the control sample size minimizing $1/n + 1/n_0$. From Equation (4.17), $n = (N - n_0)/k$. Therefore, we want to choose n_0 to minimize $k/(N - n_0) + 1/n_0$. Even though n_0 must be an integer, replace it with a variable x that can take any value, and minimize

$$\frac{k}{N-x} + \frac{1}{x} \tag{4.18}$$

with respect to x. Differentiating with respect to x and equating to 0 gives $(N-x)/x = k^{1/2}$ which, after substituting n_0 for x and using the constraint (4.17), leads to

$$n_0 = \sqrt{k}\, n. \tag{4.19}$$

That is, $\text{var}(\bar{Y}_T - \bar{Y}_C)$ is minimized (and power is maximized) when the sample size is larger in the control arm than in each other arm by a factor of the square root of the number of non-control arms. In practice, there may be interest in comparisons between active arms as well. For fixed total sample size, enlarging the control sample size compromises power for comparisons of active arms. Therefore, some trials make n_0/n larger than 1, but not as large as $k^{1/2}$, to increase power for control comparisons while maintaining reasonable power for comparisons of active arms.

It may seem paradoxical, but increasing the control sample size n_0 actually decreases the correlation between the test statistics comparing different arms to control (again assuming known common variance). The only reason these test statistics are correlated is because they share the control arm, so increasing n_0 seems like it would increase the correlation. Think instead about the fact that a larger n_0 makes $\bar{Y}_0$ closer to a constant, namely its expectation μ_0. If $\bar{Y}_0$ were exactly μ_0, the test statistics would be completely independent.

Other multi-arm designs are more popular for phase II trials to determine which interventions have the most promise for a more definitive phase III trial. Multi-arm multi-stage (MAMS) designs begin with several arms, but drop arms that do not meet a certain minimum level of benefit compared to the control (see, for example, Royston, Parmar, and Qian, 2003; Barthel, Parmar, and Royston, 2009; Magirr et al. 2012; Wason and Jaki, 2012). A simple example would be to drop any arm whose standardized z-score for comparison with control is less than 0. It might seem that dropping bad arms should require comparable multiple comparison adjustment to selecting the best treatments. However, this is not the case. The amount of statistical adjustment required to control the familywise error rate is much less when dropping arms not meeting a minimal level of efficacy.

Bayesian multi-arm designs are often used in *platform trials*, extensive duration trials comparing multiple arms to a control. Bayesian trials use a different criterion for dropping arms, often the posterior probability that an arm is best, or among the best, given data observed thus far. See Section 9.7 for a brief discussion of Bayesian monitoring of clinical trials. Bayesian platform trials are often accompanied by *response-adaptive randomization (RAR)*. See Section 5.7 for more details of RAR.

Exercises

1. A trial was conducted in heart patients with implantable cardiac defibrillators (ICDs), devices implanted in the heart that record and correct life-threatening arrhythmias.