Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Optimal weighted Bonferroni tests and their graphical extensions

Dong Xi*[1] | Yao Chen*[2]

[1] Gilead Sciences, Foster City, California, USA
[2] Novartis Pharmaceuticals Corporation, East Hanover, New Jersey, USA

Correspondence
Dong Xi, Gilead Sciences, Foster City, CA 94404, USA.
Email: xidongdxi@gmail.com

Regulatory guidelines mandate the strong control of the familywise error rate in confirmatory clinical trials with primary and secondary objectives. Bonferroni tests are one of the popular choices for multiple comparison procedures and are building blocks of more advanced procedures. It is usually of interest to find the optimal weighted Bonferroni split for multiple hypotheses. We consider two popular quantities as the optimization objectives, which are the disjunctive power and the conjunctive power. The former is the probability to reject at least one hypothesis and the latter is the probability to reject all hypotheses. We investigate the behavior of each of them as a function of different Bonferroni splits, given assumptions about the alternative hypotheses and the correlations between test statistics. Under independent tests, unique optimal Bonferroni weights exist; under dependence, optimal Bonferroni weights may not be unique based on a fine grid search. In general, we propose an optimization algorithm based on constrained non-linear optimization and multiple starting points. The proposed algorithm efficiently identifies optimal Bonferroni weights to maximize the disjunctive or conjunctive power. In addition, we apply the proposed algorithm to graphical approaches, which include many Bonferroni-based multiple comparison procedures. Utilizing the closed testing principle, we adopt a two-step approach to find optimal graphs using the disjunctive power. We also identify a class of closed test procedures that optimize the conjunctive power. We apply the proposed algorithm to a case study to illustrate the utility of optimal graphical approaches that reflect study objectives.

KEYWORDS:
Bonferroni test; Conjunctive power; Disjunctive power; Constrained non-linear optimization; Family-wise error rate; Graphical approach

1 | INTRODUCTION

Regulatory guidelines mandate the strong control of the familywise error rate (FWER) in confirmatory clinical trials with primary and secondary objectives.[1,2] Thus, the probability to reject at least one true null hypothesis incorrectly must be controlled at a pre-specified significance level $\alpha \in (0, 1)$ for any configuration of true and false null hypotheses. Bonferroni tests (unweighted and weighted), Bonferroni-based graphical approaches,[3–5] and Bonferroni-based chain procedures[6] are popular choices to control the FWER due to their simplicity to operate. This class includes fixed sequence (or hierarchical) procedures,[7] Holm procedures,[8] fallback procedures,[9,10] several gatekeeping procedures.[11–13]

*These authors contributed equally to this work.

A preferred property of a statistical test is to maximize its power while controlling its Type I error rate at a given significance level α. Thus power and sample size considerations are an integral part of clinical trial design.[14] For multiple comparisons, the conventional definition of power has been extended to deal with various success criteria based on significant outcomes from some or all hypotheses. The disjunctive power is the probability of rejecting at least one false null hypothesis.[15] This represents a natural link to the trial success criterion that requires at least one primary hypothesis being rejected. On the other hand, the conjunctive power is the probability of rejecting all false null hypotheses.[15] This reflects a stretched goal for a trial, that is, to demonstrate the superiority of the investigational treatment, an ideal case is to reject all null hypotheses. Other definitions include the average of power for each hypothesis and a weighted average with weights based on their importance.[16,17] Another related criterion is the expected number of rejected hypotheses.[18]

In the class of Bonferroni-based multiple comparison procedures, a natural question is about the optimal way to split α, i.e. to assign optimal Bonferroni weights to hypotheses. This question has been investigated for more conventional multiple comparison procedures. For example, optimal weights are provided numerically for fixed sequence procedures by maximizing the weighted power[17] and the expected number of rejections.[19] Similarly, analytical formulae are derived for weighted Bonferroni tests to obtain the optimal average power.[20,21] Examples of optimal weights using grid search and simulation are discussed for graphical approaches[22] and chain procedures.[23,24] More recently, optimization using deep learning has been proposed to optimize the weighted power for the graphical approach.[25]

Although the weighted (or average) power is easier to optimize numerically, it introduces another set of weights for power that may not be easy to determine or interpret. Thus it is commented that 'While all of the power definitions are used, perhaps "all-pairs power" and "any-pair power" are more common',[17] where "all-pairs power" is the conjunctive power and "any-pair power" is the disjunctive power. For up to three hypotheses, conjunctive and disjunctive power are explicitly derived for some multiple comparison procedures.[26] For more hypotheses or more flexible approaches such as the graphical approach, simulation over a grid search of possible-but-often-limited choices is often adopted to search for the optimal solution.[22–24] In general, there are three challenges to optimize the disjunctive and conjunctive power. The first challenge is that evaluating these objective functions requires the knowledge of multivariate distributions and hence the correlation structure among test statistics. Second it may be convoluted to express the conjunctive power as a function of p-values or test statistics. As a result, simulation is often needed which is time consuming and subject to imprecision. The third challenge is that neither the disjunctive nor the conjunctive power may uniquely determine the optimal multiple comparison procedure

To tackle these challenges, we investigate the behavior of optimal weights for weighted Bonferroni tests that maximize the disjunctive or conjunctive power using a fine grid search over all possible choices of hypotheses weights and correlations. For the disjunctive power, optimal Bonferroni weights may become non-unique and change from all non-zero quantities to zero for some hypotheses when the correlation increases. Based on these empirical findings, we tailor a constrained non-linear optimization algorithm with multiple starting points to efficiently determine optimal Bonferroni weights without simulation. We apply the proposed algorithm to graphical approaches, which include many Bonferroni-based multiple comparison procedures. Utilizing the closed testing principle,[27] we adopt a two-step approach to find optimal graphs using the disjunctive power. We also identify a class of closed test procedures that optimize the conjunctive power.

The rest of the article is organized as follows. In Section 2, we introduce the disjunctive and conjunctive power as the objective function for optimization. In Sections 3 and 4, we investigate optimal weights for weighted Bonferroni tests to maximize the disjunctive and conjunctive power, respectively using numerical optimization. In Section 5, we extend the optimization idea and derive a strategy to find optimal graphical approaches using the disjunctive power. In Section 6, we explore the property of multiple comparison procedures that maximize the conjunctive power in the closed testing framework. In Section 7, we provide concluding remarks and discussions about possible future directions.

2 | OBJECTIVES FOR OPTIMIZATION

Consider testing m null hypotheses $H_i : \theta_i \leq 0,\ i = 1, \dots, m$ against the upper alternatives, while strongly controlling the FWER at an overall significance level α. A simple and widely used procedure is the weighted Bonferroni test. Assume that hypothesis H_i receives a Bonferroni weight $0 \leq w_i \leq 1$ for $i = 1, \dots, m$, where $\sum_{i=1}^{m} w_i = 1$. Given its p-value p_i, hypothesis H_i is rejected by the weighted Bonferroni test if $p_i \leq w_i\alpha$. This procedure controls the FWER because of the weighted Bonferroni inequality and it reduces to the unweighted Bonferroni test when $w_i = 1/m$ for all $i = 1, \dots, m$.

Regarding the choice of Bonferroni weights, a larger w_i yields a larger rejection region for hypothesis H_i and makes it more likely to reject H_i. Thus one could choose Bonferroni weights based on clinical importance where more important hypotheses, which one hopes to reject more often, receive larger weights and are more likely to be rejected. Although this intuition sheds light on how optimal weights should behave intuitively, quantification of optimal weights remains an open question. As a result, one could try to find weights that maximize the likelihood of rejecting any, some or all of these m hypotheses.

Under the alternative hypothesis of H_i, we assume the test statistic $z_i = \Phi^{-1}(1-p_i)$ follows an asymptotic normal distribution with mean ξ_i and variance 1, where $\Phi^{-1}(\cdot)$ is the quantile of the standard normal distribution and ξ_i is also called the non-centrality parameter. If one could test H_i at level α without multiplicity adjustment, the power to reject H_i under the alternative space is

$$d_i = P_{\xi_i}\left(p_i \le \alpha\right) = P_{\xi_i}\left[z_i \ge \Phi^{-1}(1-\alpha)\right] = 1 - \Phi\left[\Phi^{-1}(1-\alpha) - \xi_i\right]. \tag{1}$$

In this article, we call d_i in (1) the *marginal power* to reject H_i at level α under the alternative. Thus we could establish the connection between the non-centrality parameter ξ_i and the marginal power d_i via (1) as

$$\xi_i = \Phi^{-1}(1-\alpha) - \Phi^{-1}\left(1-d_i\right). \tag{2}$$

Given that the marginal power is often readily available from statistical software, we use it as an input parameter to characterize the statistical testing behavior and use the non-centrality parameter as an intermediate quantity to facilitate the internal calculation in the following discussion.

Since each hypothesis may have a different power under its Bonferroni weight w_i, the average power has been proposed to serve as an objective function for optimization, in which every power value contributes equally. In particular, the average power for $H_1, \ldots, H_m$ is

$$\frac{1}{m}\sum_{i=1}^{m} P_{\xi_i}\left(p_i \le w_i\alpha\right). \tag{3}$$

Technically, it is convenient to optimize the average power because it only relies on the univariate marginal distribution of test statistics instead of the multivariate joint distribution. Numerical solutions have been provided using constrained nonlinear optimization to obtain weights that maximize the average power in (3).[28] Closed-form solutions were independently proposed to estimate the optimal weights.[20,21] Given this explicit formula, optimal weights that maximize the average power is popular for genetic studies but the interpretation of the average power is difficult to communicate for clinical trials. In particular, the resulting optimal weights are larger for hypotheses with mid-range non-centrality parameters and smaller for hypotheses with small or large non-centrality parameters.[21] This may not be aligned with the clinical intuition that more important hypotheses, which one hopes to reject more often, should receive larger weights and thus should be more likely to be rejected.

A further extension of the average power in (3) is a weighted average power[16]

$$\frac{1}{m}\sum_{i=1}^{m} b_i P_{\xi_i}\left(p_i \le w_i\alpha\right)$$

based on a set of pre-defined weights $\left\{b_i\right\}_{i=1}^{m}$ with $\sum_{i=1}^{m} b_i = 1$. Optimal weights are calculated using constrained nonlinear optimization[17] and using deep learning.[25] Although feasible, the weighted average power introduces another layer of complexity and further makes the interpretation even more difficult.

As in the earlier comment,[17] the disjunctive and the conjunctive power are more commonly used than the (weighted) average power. The disjunctive power is linked to the success of the whole testing strategy to reject at least one hypothesis under the alternative space. Given the p-value p_i for hypothesis H_i, the disjunctive power for the weighted Bonferroni test is

$$P_{\xi_1,\ldots,\xi_m}\left[\bigcup_{i=1}^{m}\left(p_i \le w_i\alpha\right)\right] = 1 - P_{\xi_1,\ldots,\xi_m}\left\{\bigcap_{i=1}^{m}\left[z_i < \Phi^{-1}\left(1-w_i\alpha\right)\right]\right\},$$

where the probability is evaluated under a multivariate normal distribution for $z_1, \ldots, z_m$. The mean of the distribution is $\xi_1, \ldots, \xi_m$, where $\xi_i = \Phi^{-1}(1-\alpha) - \Phi^{-1}\left(1-d_i\right)$, the variance is 1, and the correlation matrix is $R = \left\{\rho_{ij}\right\}_{i,j=1}^{m}$. Thus the disjunctive power can be further written as

$$1 - \int_{-\infty}^{\Phi^{-1}\left(1-w_1\alpha\right)-\xi_1} \cdots \int_{-\infty}^{\Phi^{-1}\left(1-w_m\alpha\right)-\xi_m} \phi\left(x_1, \ldots, x_m\right) dx_1 \cdots dx_m, \tag{4}$$

where $\phi\left(x_1, \ldots, x_m\right)$ denotes the probability density function of a multivariate normal distribution with the mean 0, the variance 1 and the correlation matrix R.

Another popular objective is the conjunctive power, which represents a stretched goal of the whole testing strategy to reject all hypotheses. Given the p-value p_i for hypothesis H_i, the conjunctive power for the weighted Bonferroni test is

$$P_{\xi_1,\ldots,\xi_m}\left[\bigcap_{i=1}^{m}\left(p_i \le w_i\alpha\right)\right] = P_{\xi_1,\ldots,\xi_m}\left\{\bigcap_{i=1}^{m}\left[z_i \ge \Phi^{-1}\left(1-w_i\alpha\right)\right]\right\},$$

where the probability is evaluated under a multivariate normal distribution with the mean $\xi_1,\ldots,\xi_m$, the variance 1, and the correlation matrix $R = \left\{\rho_{ij}\right\}_{i,j=1}^{m}$. Similarly the conjunctive power can be further written as

$$\int_{\Phi^{-1}(1-w_1\alpha)-\xi_1}^{\infty} \cdots \int_{\Phi^{-1}(1-w_m\alpha)-\xi_1}^{\infty} \phi\left(x_1,\ldots,x_m\right) dx_1 \cdots dx_m. \tag{5}$$

In the following two sections, we investigate optimal weights for the disjunctive power and then for the conjunctive power using numerical calculation. We first use a fine gird search to understand the impact of the correlations on the behavior of optimal weights and then propose a numerical optimization algorithm to efficiently find optimal weights to maximize the disjunctive and conjunctive power, respectively. R programs to reproduce calculations, tables and figures are available at https://github.com/cyustcer/multiplicity_optimal_graph.

3 | OPTIMAL BONFERRONI WEIGHTS FOR THE DISJUNCTIVE POWER

To obtain optimal weights for the maximal disjunctive power, we need to solve the following constrained optimization problem

$$\begin{aligned} \max_{w_i} \quad & \text{Disjunctive power in (4)} \\ \text{subject to} \quad & 0 \le w_i \le 1 \text{ for } i = 1,\ldots,m, \\ & \sum_{i=1}^{m} w_i = 1. \end{aligned} \tag{6}$$

From the expression in (4), the technical challenge to maximize the disjunctive power is to work with the multivariate normal distribution.

3.1 | Under independence

For simplicity, we first consider the case when all test statistics are independent and thus the correlation matrix R becomes an identity matrix. In this case, the disjunctive power reduces to $1 - \prod_{i=1}^{m} P_{\xi_i}\left(p_i > w_i\alpha\right)$, which only depends on the univariate normal distribution. The optimization problem in (6) can be formulated as

$$\begin{aligned} \max_{w_i} \quad & \left\{1 - \prod_{i=1}^{m}\Phi\left[\Phi^{-1}\left(1-w_i\alpha\right) - \Phi^{-1}(1-\alpha) + \Phi^{-1}\left(1-d_i\right)\right]\right\} \\ \text{subject to} \quad & 0 \le w_i \le 1 \text{ for } i = 1,\ldots,m, \\ & \sum_{i=1}^{m} w_i = 1. \end{aligned} \tag{7}$$

Given the set of the marginal power $\{d_i\}_{i=1}^{m}$, a unique set of Bonferroni weights $\{w_i\}_{i=1}^{m}$ exists for the optimal disjunctive power in (7) under very mild and practical conditions. For example, when the FWER is to be controlled at the one-sided level $\alpha = 0.025$, as long as the smallest marginal power $\min_i d_i$ is greater than 3%, the unique optimal solution exists. The detailed proof is provided in Appendix A.1. Based on this theoretical guarantee, we could use numerical non-linear optimization algorithms to efficiently find the unique set of optimal Bonferroni weights. In this paper, we use the Sequential Least Squares Programming (SLSQP) algorithm[29] from the `nloptr` R package.[30] A brief discussion about choices of non-linear optimization algorithms is provided in Appendix B.

To illustrate, consider three hypotheses H_1, H_2, H_3 and apply a weighted Bonferroni test to control the FWER at the one-sided level $\alpha = 0.025$. For the marginal power, we consider two scenarios. First, we assume an equal marginal power of 90% (or 80%, 70%) to reject each hypothesis. For the second scenario, we assume decreasing marginal power to reject H_1, H_2, H_3

such as (90%, 75%, 60%), (90%, 75%, 30%), (90%, 75%, 10%), (90%, 50%, 50%), and (90%, 10%, 10%), respectively. Optimal weights that maximize the disjunctive power are provided in Table 1. When the marginal power is equal for H_1, H_2, H_3, the optimal weight is also equal and always 1/3, regardless of the exact number of the marginal power. In addition, optimal weights are also equal as long as the marginal power is constant for a subset of hypotheses, e.g., for only H_2 and H_3 in the last two scenarios. When the marginal power is different for different hypotheses, the one with the higher marginal power receives a larger optimal weight and the one with the lower marginal power receives a smaller optimal weight. In all cases, the optimal disjunctive power is at least as large as the largest marginal power. Results for two or four hypotheses are provided in Appendix C and D and similar conclusions apply.

TABLE 1 Optimal Bonferroni weights for the disjunctive power with three hypotheses under independence.

Marginal power (%)	Correlation	Optimal weights			Disjunctive
d_1, d_2, d_3	ρ	w_1	w_2	w_3	power (%)
90, 90, 90	0	1/3	1/3	1/3	99.220
80, 80, 80	0	1/3	1/3	1/3	96.007
70, 70, 70	0	1/3	1/3	1/3	90.011
90, 75, 60	0	0.536	0.299	0.164	95.740
90, 75, 30	0	0.626	0.353	0.021	94.694
90, 75, 10	0	0.639	0.361	0	94.629
90, 50, 50	0	0.715	0.142	0.142	92.530
90, 10, 10	0	1	0	0	90.000

To summarize the behavior of optimal Bonferroni weights under independence for the disjunctive power, they are mainly driven by the marginal power, i.e., the higher the marginal power, the larger the optimal weight. This aligns with the clinical intuition in Section 2 very well. As a result, the hypothesis that is more likely to be rejected (without multiplicity adjustments) receives a larger weight and thus is still more likely to be rejected with the optimal Bonferroni test.

3.2 | Under dependence

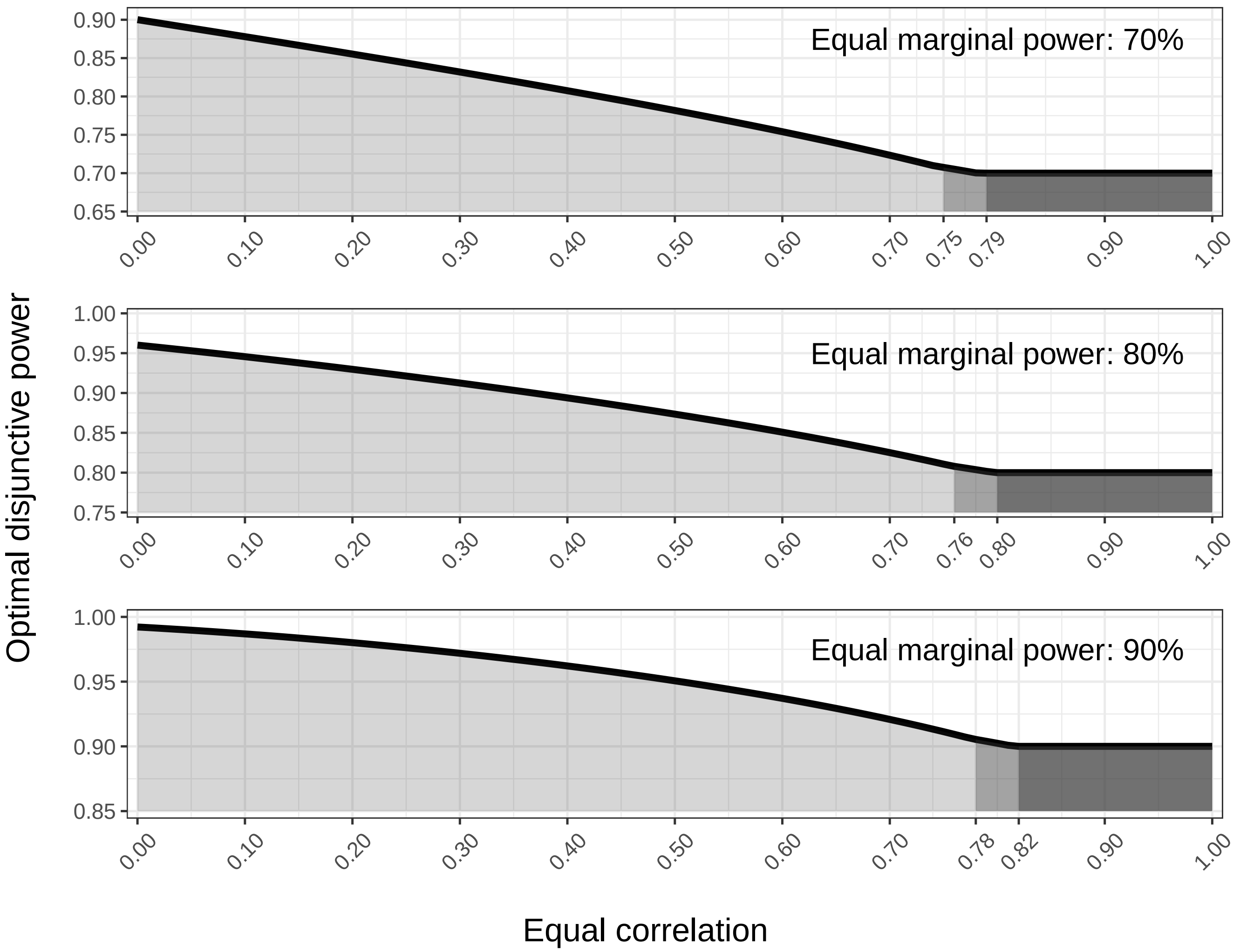
When test statistics are correlated, the optimization problem in (6) needs to be solved under a multivariate normal distribution. In addition to the marginal power, we also need to make an assumption about the correlation matrix R. In some cases, the correlation structure can be determined by design via the sample size or the number of events. Examples include the Dunnett-type[31] and the Tukey-type[32] correlation structures. In other cases, we need to hypothesize plausible correlations based on relevant data and knowledge. In many clinical trials with complex multiple comparison problems, a combination of both approaches may be applied. An example is illustrated in Section 5.3. Until then, we assume that we have reliable estimates for both the marginal power and the correlation matrix. The calculation of probabilities under multivariate normal distributions is done using the `mvtnorm` R package.[33]

For the constrained non-linear optimization in (6), the marginal power and the correlation matrix could have a joint effect on optimal weights. Given the potential complexity, we will first use a fine grid search algorithm to understand the behavior of optimal weights under simple settings with the equal marginal power and equal correlations. Then we generalize the findings with a heuristic algorithm to obtain optimal weights in general cases.

To illustrate, consider three hypotheses H_1, H_2 and H_3, and apply a weighted Bonferroni test to control the FWER at the one-sided level $\alpha = 0.025$. For the marginal power, we first assume an equal marginal power of 90% (80%, 70%) to reject each hypothesis. For the correlation matrix, we consider a compound symmetry structure with equal correlation values. For each correlation value from 0 to 1 with an increment of 0.01, optimal weights that maximize the disjunctive power are calculated based on a grid search of Bonferroni weights with an increment of 0.001. In Figure 1, we plot the behavior of the optimal disjunctive power as correlation increases. For each marginal power scenario, the optimal disjunctive power decreases as correlation increases until the optimal disjunctive power plateaus at the corresponding marginal power when the correlation is very high (e.g., well above 0.8). In addition, we plot three areas with different levels of transparency to illustrate three possible sets of optimal weights. When the correlation is low, the lightest grey area represents the case when optimal weights are $(1/3, 1/3, 1/3)$ for

H_1, H_2, H_3, respectively. In the middle area where correlation is in the neighborhood of 0.8, optimal weights are not unique and they can be $(1/2, 1/2, 0)$, $(1/2, 0, 1/2)$ or $(0, 1/2, 1/2)$ for H_1, H_2, H_3, respectively. When the correlation is very high (e.g., well above 0.8), the darkest grey area represents the case when optimal weights are not unique and they can be $(1, 0, 0)$, $(0, 1, 0)$ or $(0, 0, 1)$ for H_1, H_2, H_3, respectively. The cutoff correlation values when the structure of optimal weights changes from non-zero to zero depend on the marginal power.

FIGURE 1 Optimal disjunctive power for three hypotheses with an equal marginal power and equal correlations. Three shaded areas (from the lightest to the darkest) correspond to three sets of optimal weights: $1/3$ for each hypothesis, $1/2$ for any two hypotheses, and 1 for any one hypothesis.



More specifically, we contrast optimal weights under dependence against those obtained under independence in Table 2. When the correlation is very high (e.g., 0.9), optimal weights allocate all the weight to a single hypothesis, instead of splitting it to all hypotheses as in the independence case, which we include in the table as benchmarks. The gain in the disjunctive power ranges from 2.7% to 5.7% when the marginal power changes from 90% to 70%. When the correlation is not high (e.g., < 0.7), optimal weights are the equal Bonferroni split, which is the same as in the independence case. When the correlation is in the neighbourhood of 0.8 (e.g., 0.78), optimal weights choose two hypotheses, do the equal split between them, and leave the third hypothesis with 0 weight (e.g., 0.5, 0.5, 0). The gain in the disjunctive power is less than 1%. Although this phenomenon does not happen for a wide range of correlation, it represents a transition of optimal weights from the extreme on the boundary of the feasible region (e.g., 1, 0, 0) to the extreme in the interior (e.g., 1/3, 1/3, 1/3). This knowledge will be useful for the general optimization in Section 3.3.

Marginal power (%) H_1, H_2, H_3	Correlation ρ_{12}	ρ_{13}	ρ_{23}	Optimal weights w_1	w_2	w_3	Disjunctive power (%)
(0.9, 0.9, 0.9)	0.9	0.9	0.9	1	0	0	90.000
				0	1	0	90.000
				0	0	1	90.000
	Benchmark*			1/3	1/3	1/3	87.301
	0.78	0.78	0.78	0.5	0.5	0	90.535
				0.5	0	0.5	90.535
				0	0.5	0.5	90.535
	Benchmark			1/3	1/3	1/3	90.507
	0.7	0.7	0.7	1/3	1/3	1/3	92.082
	0.5	0.5	0.5	1/3	1/3	1/3	95.062
(0.8, 0.8, 0.8)	0.9	0.9	0.9	1	0	0	80.000
				0	1	0	80.000
				0	0	1	80.000
	Benchmark			1/3	1/3	1/3	75.502
	0.78	0.78	0.78	0.5	0.5	0	80.373
				0.5	0	0.5	80.373
				0	0.5	0.5	80.373
	Benchmark			1/3	1/3	1/3	80.130
	0.7	0.7	0.7	1/3	1/3	1/3	82.512
	0.5	0.5	0.5	1/3	1/3	1/3	87.342
(0.7, 0.7, 0.7)	0.9	0.9	0.9	1	0	0	70.000
				0	1	0	70.000
				0	0	1	70.000
	Benchmark			1/3	1/3	1/3	64.272
	0.78	0.78	0.78	0.5	0.5	0	70.034
				0.5	0	0.5	70.034
				0	0.5	0.5	70.034
	Benchmark			1/3	1/3	1/3	69.554
	0.7	0.7	0.7	1/3	1/3	1/3	72.338
	0.5	0.5	0.5	1/3	1/3	1/3	78.174

*: Different optimal weights under independence from Table 1 to illustrate the conservatism of them under dependence.

TABLE 2 Optimal Bonferroni weights to maximize the disjunctive power with three hypotheses under dependence with the equal marginal power.

Further, we consider the decreasing marginal power to reject H_1, H_2, H_3 including (90%, 75%, 60%), (90%, 75%, 30%), and (90%, 50%, 50%), respectively. For the correlation matrix, we consider both homogeneous and heterogeneous structures. Optimal weights that maximize the disjunctive power are provided in Table 3. We observe a similar trend as the independence case in Table 2: the higher the marginal power, the larger the optimal weight. As the correlation increases, this unequal split becomes more prevail and more extreme. For example, when the marginal power is (90%, 75%, 60%) for H_1, H_2, H_3 respectively, the optimal weight is (0.536, 0.299, 0.164) under independence, (0.781, 0.201, 0.018) under correlation of 0.4, and (1, 0, 0) under correlation above 0.8. When correlation increases between two hypotheses, the optimal weight increases for the hypothesis with a larger marginal power and decrease for the one with a lower marginal power. For example, when the marginal power is (90%, 75%, 60%) for H_1, H_2, H_3 respectively, the optimal weight for H_1 increases from 0.781 to 0.953 when ρ_{12} changes from 0.4 to 0.8 and the optimal weight for H_2 decreases from 0.201 to 0.000. A more visual illustration for two hypotheses is given in Figure C1 in Appendix C. The gain in the disjunctive power from optimal weights under dependence ranges from 0.3% to 4.2%, compared to that from optimal weights under independence (benchmarks).

Marginal power (%)	Correlation			Optimal weights			Disjunctive
H_1, H_2, H_3	ρ_{12}	ρ_{13}	ρ_{23}	w_1	w_2	w_3	power (%)
(90, 75, 60)	0.8	0.8	0.8	1	0	0	90.000
	Benchmark*			0.536	0.299	0.164	85.795
	0.4	0.4	0.4	0.781	0.201	0.018	91.295
	Benchmark			0.536	0.299	0.164	90.722
	0.8	0.4	0.2	0.953	0.000	0.047	90.160
	Benchmark			0.536	0.299	0.164	87.659
	0.2	0.4	0.8	0.697	0.303	0.000	92.925
	Benchmark			0.536	0.299	0.164	91.802
(90, 75, 30)	0.8	0.8	0.8	1	0	0	90.000
	Benchmark			0.626	0.353	0.021	87.092
	0.4	0.4	0.4	0.792	0.208	0.000	91.259
	Benchmark			0.626	0.353	0.021	90.939
	0.8	0.4	0.2	1	0	0	90.000
	Benchmark			0.626	0.353	0.021	87.144
	0.2	0.4	0.8	0.697	0.303	0.000	92.925
	Benchmark			0.626	0.353	0.021	92.765
(90, 50, 50)	0.8	0.8	0.8	1	0	0	90.000
	Benchmark			0.715	0.142	0.142	87.345
	0.4	0.4	0.4	0.990	0.005	0.005	90.020
	Benchmark			0.715	0.142	0.142	88.902
	0.8	0.4	0.2	0.994	0.000	0.006	90.011
	Benchmark			0.715	0.142	0.142	88.198
	0.2	0.4	0.8	0.920	0.080	0.000	90.421
	Benchmark			0.715	0.142	0.142	89.319

*: Different optimal weights under independence from Table 1 to illustrate the conservatism of them under dependence.

TABLE 3 Optimal Bonferroni weights to maximize the disjunctive power with three hypotheses under dependence with unequal values of the marginal power.

To summarize the behavior of optimal weights under dependence, when the marginal power is constant, optimal weights are always equally split among non-zero-weighted hypotheses. And the higher the correlation, the fewer non-zero-weighted hypotheses and the more zero-weighted hypotheses. For example, when the correlation is not high, all m hypotheses receive an optimal weight of $1/m$; when the correlation increases, the optimal strategy selects $m-1$ out of m hypotheses and assigns $1/(m-1)$ weight to each of them (all $\binom{m}{m-1}$ selections lead to the same disjunctive power); when the correlation keeps increasing, there are fewer non-zero-weighted hypotheses, among which optimal weights are equal; finally when the correlation is very high, one hypothesis gets the optimal weight of 1. The role of the correlation is to push optimal weights to extremes–when the correlation increases between hypotheses, the hypothesis with a higher marginal power receives an increased optimal weight while the one with a lower marginal power receives a decreased optimal weight.

3.3 | Heuristic algorithm using numerical optimization

Although the grid search gave us a comprehensive solution of optimal weights, it is computationally intensive and not scalable as the number of hypotheses increases. As a general solution, we adopt the SLSQP algorithm in the R package `nloptr`[30] and propose a heuristic strategy based on the pattern of optimal weights we learned from the grid search.

From the investigation in Section 3.2, the optimization of the disjunctive power might be a non-convex optimization problem depending on the correlation (see Figure C1 in Appendix C for the case with two hypotheses). One important aspect of non-convex optimization is about the choice of starting points. Based on these empirical findings from the grid search, sub-optimal weights could be found and incorrectly considered as the optimal solution, because the starting point is far away from the global optimal. In these cases, the algorithm is trapped around the local optima even we increase the number of numerical searches. To alleviate this unsatisfactory performance, we transfer our learning from the grid search with three hypotheses in Section 3.2. Thus we choose the starting points to cover possible optimal weights in the case of the equal marginal power and equal correlations. In particular, for m hypotheses $H_1, \dots, H_m$, we consider the following starting points (iteration=0) for $k = 1, \dots, 2^m - 1$,

$$\boldsymbol{w}^{k(0)} = \begin{cases} (1/m, \dots, 1/m), \\ \text{equal weights of } 1/(m-1) \text{ for all } \binom{m}{m-1} \text{ sets of } m-1 \text{ hypotheses}, \\ \text{equal weights of } 1/(m-2) \text{ for all } \binom{m}{m-2} \text{ sets of } m-2 \text{ hypotheses}, \\ \dots \\ \text{weight of 1 to each of } \binom{m}{1} \text{ hypotheses}. \end{cases} \tag{8}$$

For each set of starting points, we save the local optimal weights from the SLSQP algorithm. Finally, the overall optimal weights are provided as the set that leads to the maximal disjunctive power among the above $2^m - 1$ cases. This process is provided in Algorithm 1.

Algorithm 1 Optimal Bonferroni weights for the disjunctive power

```
Initialize starting points w^{k(0)}, k = 1, ..., 2^m − 1 from (8)
for k = 1, ..., 2^m − 1 do
    Find the local optimal weights w^{k*} to maximize the disjunctive power with the given starting point w^{k(0)} in (8)
end for
Store and sort all local optimal weights by the optimal disjunctive power for all k = 1, ..., 2^m − 1
```

Although the number of starting points increases exponentially in the number of hypotheses m, the requirement for computing and memory is still much less compared to a grid search. For example, with three hypotheses considered in Section 3.2, we need to consider $2^3 - 1 = 7$ sets of starting points: $(1/3, 1/3, 1/3)$, $(1/2, 1/2, 0)$, $(1/2, 0, 1/2)$, $(0, 1/2, 1/2)$, $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$. Our proposed numerical algorithm successfully reproduce Tables 2 and 3 that the same set of optimal weights are found as the grid search. However, for each set of the marginal power and correlations, the computing time is much less than that of the grid search, because the latter has to evaluate the disjunctive power for all possible combinations of $w_i, i = 1, 2, 3$ in the increment of 0.001 such that $\sum_i w_i = 1$. Our proposed numerical algorithm only needs to consider $2^3 - 1 = 7$ sets of starting points and often finishes in a few seconds. When the number of hypotheses is more than three, the possible combinations of w_i could easily exceed the memory limit, making the grid search infeasible. But our numerical algorithm is still able to find optimal weights, which are in a similar fashion as for the case with three hypotheses, in an acceptable time frame. More evaluation of computing time is provided in Appendix B.

4 | OPTIMAL BONFERRONI WEIGHTS FOR THE CONJUNCTIVE POWER

To obtain the optimal weights for the maximal conjunctive power, we need to solve the following constrained optimization problem

$$\begin{aligned} \max_{w_i} \quad & \text{Conjunctive power in (5)} \\ \text{subject to} \quad & 0 \le w_i \le 1 \text{ for } i = 1, \dots, m, \\ & \sum_{i=1}^{m} w_i = 1 \end{aligned} \tag{9}$$

4.1 | Under independence

Similar to the optimal Bonferroni weights for the disjunctive power, we first consider the case when all test statistics are independent and thus the correlation matrix R is an identity matrix. In this case, the conjunctive power is $\prod_{i=1}^{m} P_{\xi_i}\left(p_i \le w_i \alpha\right)$ and the optimization problem in (9) can be formulated as

$$\begin{aligned} \max_{w_i} \quad & \left\{ \prod_{i=1}^{m} \Phi\left[\Phi^{-1}(1-\alpha) - \Phi^{-1}\left(1-d_i\right) - \Phi^{-1}\left(1-w_i\alpha\right)\right] \right\} \\ \text{subject to} \quad & 0 < w_i \leq 1 \text{ for } i = 1, \dots, m, \\ & \sum_{i=1}^{m} w_i = 1 \end{aligned} \tag{10}$$

Note that we exclude the possibility of $w_i = 0,\ i = 1, \dots, m$ in the constraints. This means H_i would be tested at level 0 and thus it would be impossible to reject it, leading to a conjunctive power of 0. As seen in Section 3, this boundary condition complicates the search of optimal weights for the disjunctive power. Thus excluding the zero weight simplifies the search for the conjunctive power. Given the set of the marginal power $\{d_i\}_{i=1}^m$, a unique set of Bonferroni weights $\{w_i\}_{i=1}^m$ exists for the optimal conjunctive power in (10) under very mild and practical conditions. For example, when the FWER is to be controlled at the one-sided level $\alpha = 0.025$, as long as the smallest marginal power $\min_i w_i$ is greater than 3%, the unique optimal solution exists. The detailed proof is provided in Appendix A.2. Based on this theoretical guarantee, we could use the SLSQP algorithm in the R package `nloptr`[30] to find the unique set of optimal Bonferroni weights.

To illustrate, we consider three hypotheses H_1, H_2, H_3 with both equal and unequal marginal power, similar to cases for the optimal disjunctive power. Optimal weights that maximize the conjunctive power are provided in Table 4. When the marginal power are equal across H_1, H_2, H_3, the optimal weights are always 1/3, regardless of the exact number of the marginal power. The optimal weights are also equal as long as the marginal power is constant for a subset of hypotheses. These behaviors are the same for the optimal weights maximizing the disjunctive power. However, when the marginal power is different for different hypotheses, the one with the higher marginal power receives a lower optimal weight and the one with the lower marginal power receives a higher optimal weight. Although a larger weight means a higher power for an individual hypothesis, to declare success under the conjunctive power, the balance of power for all hypothesis is more important. This behavior is opposite to the behavior of optimal weights for the disjunctive power. In particular, the optimal weights for the conjunctive power do not allow the zero weight as it will yield a conjunctive power of 0. In all cases, the optimal conjunctive power is smaller than the smallest marginal power, which reflects the more challenging nature of the conjunctive power. Results for two or four hypotheses are provided in Appendix C and D and similar conclusions apply.

Marginal power (%)	Correlation	Weights			Conjunctive
H_1, H_2, H_3	ρ	w_1	w_2	w_3	power (%)
90, 90, 90	0	1/3	1/3	1/3	51.518
80, 80, 80	0	1/3	1/3	1/3	28.517
70, 70, 70	0	1/3	1/3	1/3	15.400
90, 75, 60	0	0.208	0.338	0.454	21.268
90, 75, 30	0	0.167	0.268	0.565	9.194
90, 75, 10	0	0.137	0.219	0.644	2.706
90, 50, 50	0	0.166	0.417	0.417	9.566
90, 10, 10	0	0.102	0.449	0.449	0.198

TABLE 4 Optimal Bonferroni weights for the conjunctive power with three hypotheses under independence.

4.2 | Under dependent tests

When test statistics are correlated, the optimization problem in (9) needs to be solved under a multivariate normal distribution to obtain optimal weights. Similar to our approach to the disjunctive power, we will first use the grid search to learn the behavior of optimal weights under simple settings with equal marginal power and equal correlations. Then we generalize the findings using numerical optimization to obtain optimal weights in general cases.

To illustrate, consider three hypotheses H_1, H_2, H_3 and apply a weighted Bonferroni test to control the FWER at one-sided level $\alpha = 0.025$. For the marginal power, we first assume an equal marginal power of 90% (80%, 70%) to reject each hypothesis. For the correlation matrix, we consider a compound symmetry structure with equal correlation values. For each correlation from 0 to 1 with an increment of 0.01, optimal weights that maximize disjunctive power are calculated based on a grid search of an increment of 0.001. Contrary to the complex results for the disjunctive power, optimal weights for the conjunctive power are always $(1/3, 1/3, 1/3)$ for H_1, H_2, H_3, respectively, regardless of the value of equal marginal power and equal correlations.

Further, we consider the decreasing marginal power to reject H_1, H_2, H_3 including (90%, 75%, 60%), (90%, 75%, 30%), and (90%, 50%, 50%), respectively. For the correlation matrix, we consider both homogeneous and heterogeneous structures. Optimal weights that maximize the conjunctive power are provided in Table 5. We observe a similar trend as the independence case in Table 4: the higher the marginal power, the smaller the optimal weight. As the correlation increases, this unequal split becomes more prevail and more extreme. For example, when the marginal power is (90%, 75%, 60%) for H_1, H_2, H_3 respectively, the optimal weight is (0.208, 0.338, 0.454) under independence, (0.172, 0.331, 0.197) under correlation of 0.4, and (0.121, 0.314, 0.565) under correlation of 0.8. The gain in the conjunctive power from optimal weights under dependence ranges from 0.02% to 2.27%, compared to that from optimal weights under independence (benchmarks).

Marginal power (%)	Correlation			Optimal weights			Conjunctive
H_1, H_2, H_3	ρ_{12}	ρ_{13}	ρ_{23}	w_1	w_2	w_3	power (%)
(90, 75, 60)	0.8	0.8	0.8	0.121	0.314	0.565	42.451
	Benchmark*			0.208	0.338	0.454	41.520
	0.4	0.4	0.4	0.172	0.331	0.497	31.079
	Benchmark			0.208	0.338	0.454	30.965
	0.8	0.4	0.2	0.109	0.347	0.544	31.577
	Benchmark			0.208	0.338	0.454	30.688
	0.2	0.4	0.8	0.225	0.295	0.480	36.184
	Benchmark			0.208	0.338	0.454	36.074
(90, 75, 30)	0.8	0.8	0.8	0.055	0.153	0.792	23.769
	Benchmark			0.167	0.268	0.565	21.508
	0.4	0.4	0.4	0.115	0.225	0.660	16.088
	Benchmark			0.167	0.268	0.565	15.806
	0.8	0.4	0.2	0.066	0.253	0.681	15.525
	Benchmark			0.167	0.268	0.565	14.771
	0.2	0.4	0.8	0.150	0.149	0.701	20.049
	Benchmark			0.167	0.268	0.565	19.144
(90, 50, 50)	0.8	0.8	0.8	0.064	0.468	0.468	27.486
	Benchmark*			0.166	0.417	0.417	26.528
	0.4	0.4	0.4	0.118	0.441	0.441	17.576
	Benchmark			0.166	0.417	0.417	17.435
	0.8	0.4	0.2	0.050	0.464	0.486	16.986
	Benchmark			0.166	0.417	0.417	15.952
	0.2	0.4	0.8	0.174	0.432	0.394	22.719
	Benchmark			0.166	0.417	0.417	22.695

*: Different optimal weights under independence from Table 4 to illustrate the conservatism of the Bonferroni split under dependence.

TABLE 5 Optimal Bonferroni weights to maximize the conjunctive power with three hypotheses under dependence with the unequal marginal power.

To summarize the behavior of optimal weights for the conjunctive power, when the marginal power is equal, optimal weights are always equally split among hypotheses, if the correlation is also equal between any pair of two hypotheses. When the marginal power is different, the optimal weight is smaller for a hypothesis with a higher marginal power and is larger for a hypothesis with a lower marginal power. The role of the correlation is to push optimal weights to extremes–when the correlation increases between hypotheses, the hypothesis with a higher marginal weight has a decreased optimal weight while the one with a lower marginal power has an increased optimal weight. This behavior of optimal weights aligns intuitively with the definition of the conjunctive power because it is bounded above by the smallest marginal power and thus by increasing weights for the hypothesis with the smallest marginal power will potentially increase the conjunctive power.

4.3 | General solution via numerical optimization

As a general solution, we adopt the numerical optimization algorithm SLSQP algorithm in the R package `nloptr`.[30] Because optimal weights for the conjunctive power are always in the interior of the feasible region, we only consider one set of starting points of equal weights. For three hypotheses, we reproduce all results previously using grid search in this section, in particular Tables 4 and 5.

5 | OPTIMAL GRAPHICAL APPROACHES USING THE DISJUNCTIVE POWER

As a general framework, graphical approaches provide flexibility and transparency in designing, visualizing, and performing FWER-controlling multiple test procedures.[3,4] In this framework, null hypotheses are represented by nodes with weights associated with their local significance levels. Weighted directed edges between nodes specify the propagation of the local significance levels, that is, when a null hypothesis is rejected, its local weight is split and propagated to other not-yet-rejected hypotheses.

Consider testing m elementary null hypotheses $H_i, i = 1, \ldots, m$, while strongly controlling the FWER at an overall significance level α. Using the graphical approach,[3] we assign the hypothesis weight w_i for each $H_i, i = 1, \ldots, m$. In addition, we denote by g_{ij} the transition weight for the edge from H_i to H_j. Let p_i denote the p-value for $H_i, i = 1, \ldots, m$. When $p_i \leq \alpha w_i$, H_i can be rejected. Following the propagation, the fraction g_{ij} of its local weight w_i will be transferred to H_j so that H_j's hypothesis weight becomes $w_j + w_i g_{ij}$. Then we could compare p_j with this updated significance level $\alpha(w_j + w_i g_{ij})$. Formally, an algorithm to update hypothesis weights and transition weights is provided.[3] It has been proved that this graphical multiple comparison procedure is equivalent to a closed test procedure[27] using the α-level weighted Bonferroni test for each intersection hypothesis and thus it controls the FWER strongly at level α. Therefore, the unknown parameters in the graphical approach are hypothesis weights $\boldsymbol{w} = \left\{w_i, i = 1, \ldots, m\right\}$ and transition weights $\boldsymbol{g} = \left\{g_{ij}, i, j = 1, \ldots, m\right\}$, which we will provide optimal strategies to select, subject to the regularity conditions

$$0 \leq w_i \leq 1, \ \sum_{i=1}^{m} w_i = 1, \ 0 \leq g_{ij} \leq 1, g_{ii} = 0, \text{ and } \sum_{j=1}^{m} g_{ij} = 1. \tag{11}$$

Note that here we require equality (instead of $\leq$) in $\sum_{i=1}^{m} w_i = 1$ and $\sum_{j=1}^{m} g_{ij} = 1$ because this leads to uniformly more powerful procedures.[3]

5.1 | Motivating example

As an illustrative example, we consider a 24-week multi-center, randomized, double-blind, placebo controlled study to compare an investigational treatment against placebo with the purpose to reduce the relapse rate during the 24 weeks of study.[22] There are three doses of the investigational treatment to be compared with placebo control. The primary endpoint is the difference in proportions of patients with recurrence of relapses at week 24. The secondary endpoint is the difference in means of change in the total medication score from baseline to week 24. The three dose-control comparisons for the primary endpoint generate three hypotheses $H_i, i = 1, \ldots, 3$ for the high, medium and low dose versus control, respective. There are three corresponding secondary hypotheses $H_i, i = 4, \ldots, 6$ for three doses versus control, respectively.

The question of interest is to find the optimal graph with respect to hypothesis weights and transition weights using the knowledge of optimal Bonferroni weights for the disjunctive power. There is one clinical consideration: A secondary hypothesis is not considered unless its primary hypothesis for the same dose-control comparison has been rejected. This consideration can be translated to the following constraints on hypothesis weights and transition weights:

- Initial hypothesis weights for secondary hypotheses should be 0.

- The transition weight from a primary hypothesis to the secondary hypothesis for another dose-control comparison should be 0.
- The transition weight from a secondary hypothesis to another secondary hypothesis should be 0.
- The transition weight from a secondary hypothesis to its parent primary hypothesis for the same dose-control comparison should be 0.

To reflect these constraints above, we provide a general graph in Figure 2, where remaining unspecified hypothesis weights and transition weights are all parameters to be optimized over subject to the regularity conditions in (11). Therefore, we want to find optimal $0 \leq w_1, w_2, w_3 \leq 1$ subject to $\sum_{i=1}^{3} w_i = 1$, and optimal transition weights subject to $\sum_{j=1}^{6} g_{ij} = 1$ for $i = 1, \dots, 6$ with constraints listed above. The transition matrix incorporated with constraints is expressed in (12).

$$\begin{bmatrix} 0 & g_{12} & g_{13} & g_{14} & 0 & 0 \\ g_{21} & 0 & g_{23} & 0 & g_{25} & 0 \\ g_{31} & g_{32} & 0 & 0 & g_{36} & 0 \\ 0 & g_{42} & g_{43} & 0 & 0 & 0 \\ g_{51} & 0 & g_{53} & 0 & 0 & 0 \\ g_{61} & g_{62} & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{12}$$

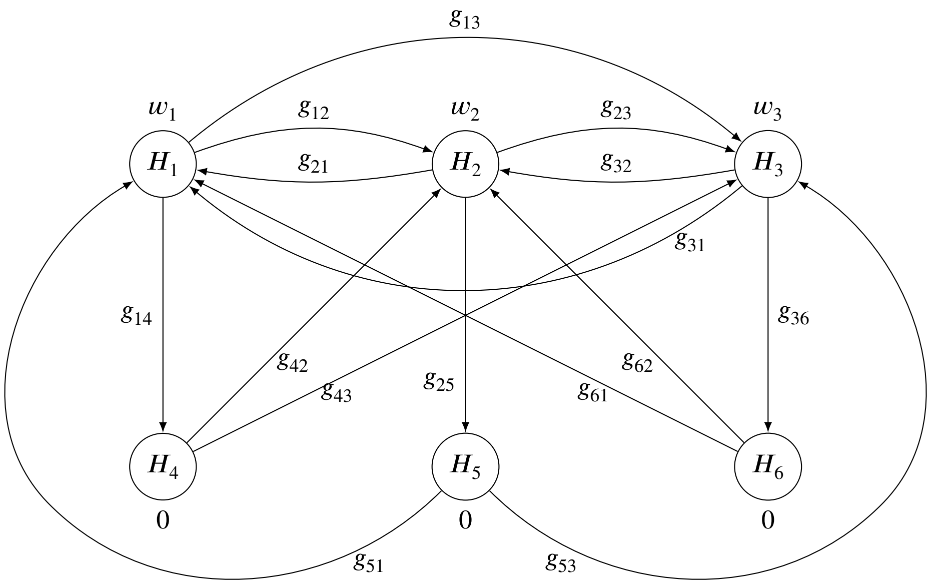


FIGURE 2 A general graph for the case study.

5.2 | Optimization with constraints

From the motivating example in Section 5.1, we can see the challenges to find optimal graphs using the disjunctive power. First, clinical and other considerations may lead to a priori constraints on hypothesis weights and transition weights. On one hand, these may reduce the dimensionality of the parameter space, which is at the quadratic order of the number of hypotheses; on the other hand, we need to modify the feasible region to accommodate these constraints. Second, optimizing the disjunctive power may not be able to uniquely determine a graph. As seen from Section 3, optimizing the disjunctive power once only determines hypothesis weights with no impact on transition weights. This is because the disjunctive power is only concerned with the first rejected hypothesis but is not affected by further rejections; transition weights only affect further rejections. Thus we need to adopt a two-step optimization using the disjunctive power as a tool to uniquely determine a graph.

For hypothesis weights $\boldsymbol{w} = \left\{w_i, i = 1, \dots, m\right\}$, optimization is the same as optimizing Bonferroni weights because the first step of the graphical testing is to reject any H_i if $p_i \leq \alpha w_i$. Thus the same technique in Section 3 can be applied with possibly additional constraints, e.g., $w_i = 0$ for some i. This can be achieved by excluding these hypotheses from the optimization and subtracting their hypothesis weights from the optimization's constraint.

For transition weights $\boldsymbol{g} = \left\{g_{ij}, i, j = 1, \dots, m\right\}$, they can be optimized as a result of optimizing Bonferroni weights for the remaining $m-1$ hypotheses after removing each hypothesis from the graph. For example, after we obtain the optimal hypothesis weights $w_i, i = 1, \dots, m$ for the disjunctive power, we want to determine transition weights g_{ji} for edges originating from H_j. So we first assume that H_j has been rejected and thus removed from the graph. The updated hypothesis weights for the remaining hypotheses H_i are $w_i^* = w_i + w_j g_{ji}, i \in \{1, \dots, m\}/\{j\}$, following the updating algorithm of the graphical approaches.[3] On the other hand, we could apply the optimization algorithm to the remaining hypotheses to find optimal Bonferroni weights for the disjunctive power. Additional constraints of $w_i' \geq w_i, i = 2 \dots, m$ need to be applied as we assumed that $g_{ij} \geq 0$ for all $i,\ j$. Some w_i' may be set to be a specific number because g_{ji} is determined by a priori constraints, e.g., $w_i' = w_i$ because of $g_{ji} = 0$ a priori. Once we get the set of optimal weights w_i' to optimize the disjunctive power of $H_i, i \in \{1, \dots, m\} \setminus \{j\}$ under constraints, we equate $w_i^* = w_i'$. Thus the transition weight for the edge from H_j to H_i can be obtained as $\left(w_i' - w_i\right)/w_j$. More formally, the optimization strategy is as follows:

Algorithm 2 Optimal graphical approaches using the disjunctive power

1: For hypotheses $H_i, i = 1, \dots, m$, the hypothesis weight w_i and the transition weight g_{ij} should satisfy the regularity conditions in (11)
2: Incorporate additional a priori constraints following clinical and other considerations, e.g., $w_i = 0$ or $g_{ij} = 0$ for some $i, j \in \{i = 1, \dots, m\}$
3: Obtain the optimal hypothesis weights $w_i, i = 1, \dots, m$ for the disjunctive power using Algorithm 1
4: **for** $j = 1, \dots, m$ **do**
5: If $w_j = 0$, temporarily assign $w_j = \varepsilon > 0$ and normalize the rest of hypothesis weights, where ε is a small value.
6: Obtain the optimal hypothesis weights $w_i', i \in \{1, \dots, m\} \setminus \{j\}$ for the disjunctive power using Algorithm 1 and additional constraints $w_i' \geq w_i$
7: Obtain the optimal transition weights $g_{ji} = \left(w_i' - w_i\right)/w_j$
8: **end for**
9: Optimal hypothesis weights $\boldsymbol{w} = \left\{w_i, i = 1, \dots, m\right\}$ and optimal transitions $\boldsymbol{g} = \left\{g_{ij}, i, j = 1, \dots, m\right\}$ are obtained using the disjunctive power

One technical point in Step 5 of Algorithm 2 is that $g_{ji} = \left(w_i' - w_i\right)/w_j$ may not be properly defined if $w_j = 0$, which could happen for secondary hypotheses, e.g., in the motivating example in Section 5.1. We provide a practical solution by assigning a very small weights to $w_j = \varepsilon$ and normalizing the rest of hypothesis weights proportionally to be $w_i(1-\varepsilon)$ so that the summation of all hypothesis weights is still 1. Note that this modification is only to overcome the technical issue in Step 5 and it should not change the optimal hypotheses weights in Step 3. More technical discussions about this practical solution are provided in Appendix E.

5.3 | Motivating example revisited

In the motivating example in Section 5.1, there are three doses of treatment compared with placebo control on one primary endpoint and one secondary endpoint. These lead to three primary hypotheses H_1, H_2, H_3 and three secondary hypotheses H_4, H_5, H_6, for three dose-control comparisons respectively. The question of interest is to find the optimal graph with respect to hypothesis weights and transition weights subject to the following clinical consideration: A secondary hypothesis is not considered unless its primary hypothesis for the same dose-control comparison has been rejected. This consideration translates to $w_4 = w_5 = w_6 = 0$ and the transition matrix with constraints in (12).

For the primary endpoint on the proportion of patients with recurrence of relapses, the lower the proportion, the more favorable the outcome. Assume the proportions are 0.181 for dose 1, 2, 3 and 0.3 for control. Using an equal randomization among the four treatment groups, the clinical trial team choose a total sample size of 800 with 200 per treatment group. This leads to a marginal power of 80% for each of dose 1, 2, 3 versus placebo, using the two-sample test for difference in proportions with unpooled variance each at one-sided $\alpha = 0.025$ significance level.

For the secondary endpoint of change in total medication score from baseline, the more reduction, the more favorable the outcome. Assume the mean change from baseline is the reduction of 11.5, 10.6, 10 and 5, respectively for dose 1, 2, 3 and control. Further assume a known common standard deviation of 10. Given the total sample size of 800 with 200 per treatment group, the marginal power is 90%, 80% and 71% for dose 1, 2, 3 versus control respectively, using the two-sample z-test for the difference in means each at level one-sided $\alpha = 0.025$.

To derive the correlation matrix for test statistics, we have two groups of correlations. The correlation between test statistics for the same endpoint is determined by the sample size. For example, the correlation between test statistics for H_1 and H_2 is $\rho_{12} = \left(\frac{n_1}{n_1+n_0}\right)^{1/2}\left(\frac{n_2}{n_2+n_0}\right)^{1/2}$, where n_0, n_1, n_2 are the sample size for control, dose 1 and dose 2, respectively. This correlation reduces to 0.5 for the equal randomization setting. Similarly, we could calculate the correlation between test statistics for primary hypotheses and for secondary hypotheses separately. The second group of correlation is the correlation between endpoints, which is usually unknown at the design stage. Thus we treat it as a nuisance parameter $\rho = \rho_{13} = \rho_{24} = \rho_{36}$. Finally, for the correlation between a primary hypothesis for one dose and a secondary hypothesis for another dose, we use the product rule. For example, $\rho_{14} = \rho_{13}\rho_{34} = 0.5\rho$. Therefore, the correlation matrix $\boldsymbol{\rho}$ is available up to the nuisance parameter ρ for the correlation between endpoints. For the following calculation, we assume $\rho = 0.5$.

Now we are ready to find optimal hypothesis and transition weights following Algorithm 2. For hypothesis weights w_1, w_2, w_3 with marginal power of 80% for each of H_1, H_2, H_3, the optimal allocation is $1/3, 1/3, 1/3$ as in Table 2. To identify the optimal transition weights from H_1, we remove H_1 from the graph. The updated hypothesis weights for $H_2, \dots, H_6$ are $w_2 + g_{12}w_1$, $w_3 + g_{13}w_1$, $g_{14}w_1$, 0, 0, respectively. Thus we reapply Algorithm 1 to find optimal hypothesis weights for H_2, H_3 and H_4 with the updated lower bound of $w_2 = 1/3, w_3 = 1/3, w_4 = 0$, and optimal weights are $w_2' = 1/3, w_3' = 1/3, w_4' = 1/3$, respectively. Thus we could back calculate transitions weights, e.g., $g_{12} = (w_2' - w_2)/w_1 = 0$, $g_{13} = 0$ and $g_{14} = (w_4' - w_4)/w_1 = 1$. Similarly, we could find optimal transition weights for edges going out of H_2 and H_3.

To identify the optimal transition weights from H_4, we set $w_4 = \varepsilon = 0.001$ to avoid the zero hypothesis weight on the denominator. Then normalizing the rest of local weights gives us $w_1 = w_2 = w_3 = 1/3 - \varepsilon/3 = 0.333$. After removing H_4 from the graph. The updated local weights for H_1, H_2, H_3, H_5, H_6 are $w_1, w_2 + g_{42}w_4, w_3 + g_{43}w_4$, 0, 0, respectively. Thus we reapply Algorithm 1 to find optimal hypothesis weights for H_2 and H_3 with the updated lower bound of $w_2 = w_3 = 1/3 - \varepsilon/3$, which are $w_2' = 0.3335, w_3' = 0.3335$, respectively. Thus we could back calculate transitions weights, e.g., $g_{42} = (w_2' - w_2)/w_4 = 0.5$ and $g_{43} = 0.5$. Similarly, we could find optimal transition weights for edges going out of H_5 and H_6. The final optimal graph is illustrated in Figure 3.

Finally, we provide some intuitive explanations behind the optimal graph, especially with respect to transition weights. First, we can see that both H_1 and H_2 have only one outgoing edge to its descendant secondary hypothesis with a transition weight of 1. The reason is that each of their secondary hypotheses have a high marginal power compared with the marginal power values on the primary hypotheses for other two dose-control comparisons. For example, H_4 has a marginal power of 0.9, which is larger than the one for H_2 and H_3. Then after H_1 is rejected, the optimal hypothesis weight for H_4 should be larger compared with those for H_2 and H_3. However, H_2 and H_3 retain the local weights of $1/3$ from the first step, which means that all the weight w_1 should be propagated to H_4. Thus, $g_{14} = 1$ and $g_{12} = g_{13} = 0$. Second, there are three outgoing edges from H_3 to H_1, H_2, H_6. The reason is that H_6 has a smaller marginal power of 71% compared to the one of H_1 and H_2. Then after H_3 is rejected, the optimal hypothesis weight for H_1, H_2, H_6 should give a larger weight to H_1 and H_2 and a smaller weight to H_6. Given that $w_6 = 0$, the majority of w_3 should be propagated to H_6 and thus $g_{36} = 0.886$, which leads to $g_{31} = g_{32} = 0.057$, sharing the rest equally. Finally, each secondary hypothesis has two outgoing edges to primary hypotheses for the other two dose-control comparisons. The transition weight is 0.5 for each edge because primary hypotheses have the equal marginal power and thus the optimal hypothesis split should also be equal.

In practice, it would also be useful to check the sensitivity of the assumed correlation between endpoints. If we change from $\rho = 0.5$ as in above to $\rho = 0$, then $g_{31} = g_{32} = 0$ and $g_{36} = 1$. This time H_6 gets all the weights from H_3 even it has smaller marginal power, because the correlations between H_6 and H_1, H_2 are smaller than the correlation between H_1 and H_2, in such case the optimal hypothesis weights leaning towards H_6 rather than H_1 and H_2.

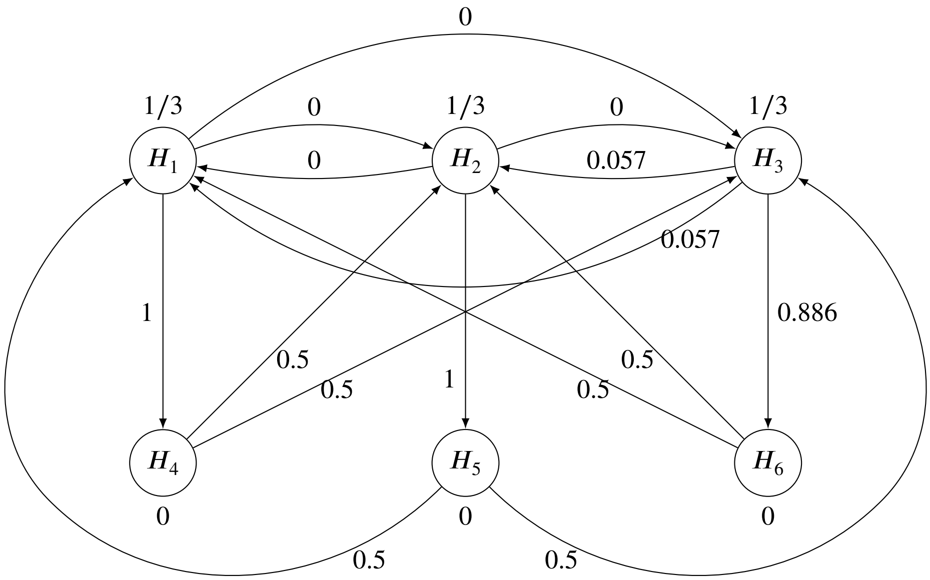


FIGURE 3 Optimal graph for the case study.

6 | OPTIMAL GRAPHICAL APPROACHES USING THE CONJUNCTIVE POWER

Because the conjunctive power is the probability to reject all hypotheses, we need to explore the configuration under which the optimal graph could achieve this goal. Recall that the graphical approaches control the FWER strongly at level α because it defines a closed test procedure using the α-level weighted Bonferroni test for each intersection hypothesis. More specifically, consider testing m elementary hypotheses H_i for $i \in I = \{1, \dots, m\}$. The closure family includes all intersection hypotheses $H_J = \bigcap_{j \in J}\{H_j\}, J \subseteq I$. An elementary hypothesis H_i is rejected by the closed test if all intersection hypotheses associated with it are rejected, i.e., all H_J with $i \in J$ are rejected. Therefore, to reject all hypotheses for the conjunctive power, the closed test is required to reject all intersection hypotheses.

We could modify the conjunctive power definition for the closure family as

$$P_{\xi_1,\dots,\xi_m}\left[\bigcap_{J \subseteq I}\left(H_J \text{ is rejected at level } \alpha\right)\right],$$

where ξ_i is the non-centrality parameter for $H_i, i \in I$. From this expression, we can derive an upper bound of the conjunctive power as

$$P_{\xi_1,\dots,\xi_m}\left[\bigcap_{i \in I}\left(H_i \text{ is rejected at level } \alpha\right)\right]. \tag{13}$$

If we can show that this upper bound is achievable, then this means that a closed test procedure is optimal for the conjunctive power if and only if it rejects all hypotheses when each hypothesis can be rejected at level α. This conclusion is applicable to all closed test procedures, which include the graphical approaches.

This upper bound of the conjunctive power in (13) is indeed achievable by commonly used closed test procedures. For example, the fixed sequence procedure uses a pre-defined order for testing, e.g., $H_1, H_2, \dots, H_m$.[7,17] Each hypothesis can be rejected at level α, as long as its p-value is less than or equal to α and all previous hypotheses are rejected. In other words, all hypotheses can be rejected when all p-values are not larger than α. However, the fixed sequence procedure that satisfies this requirement is not unique. To further determine the order of testing, one could rank hypotheses using their marginal power in a decreasing fashion or optimization under another objective, e.g., the expected number of rejected hypotheses.[19]

Another example is outside the class of Bonferroni-based closed test procedures. Simes test[34] is another popular test that can be applied to intersection hypotheses. Its rejection decisions are based on the ordered p-values and it rejects an intersection hypothesis H_J if $p_{(i)} \leq i\alpha/|J|$ for at least one $i \in J$, where $p_{(1)} \leq \cdots \leq p_{(|J|)}$ are the ordered p-values and $|J|$ is the number of elements in J. There, we observe that all intersection hypotheses are rejected if $p_{(m)} \leq \alpha$ and a closed test based on the

Simes test achieves the upper bound of the conjunctive power in (13). Therefore, the following procedures share the same optimal conjunctive power including Hommel procedures,[35] Hochberg procedures,[36] hybrid Hommel-Hochberg procedures,[37] their extensions using symmetric graphs[38] and graphical approaches[39] based on weighted Simes tests.[16] One limitation of Simes-based procedures is that they need to satisfy certain dependence conditions of the test statistics.[40–42]

7 | CONCLUSIONS AND DISCUSSIONS

Although it is widely known that Bonferroni tests are conservative under a high correlation, optimization of these weighted tests further reveal the behavior of optimal weights when the correlation increases. This has a direct implication for practical applications where the correlation may be high. For example, in the case of simultaneously testing a hypothesis in the overall population and a subpopulation, the correlation is mainly driven by the prevalence of the subpopulation, which could be larger than 0.8 if the subpopulation is more than 70% of the overall population ($\sqrt{0.7} = 0.84$). In this case of a high correlation, a Bonferroni split may not be optimal for the disjunctive power and it may be more powerful to choose a fixed sequence procedure.

In practice, how to optimize a graphical multiple comparison procedure is a frequently asked question by both statisticians and non-statisticians with the hope to increase the chance of the trial success. We adopted two commonly used power objectives, the disjunctive and conjunctive power, and used them to optimize graphs. The resulting optimal graphs using the disjunctive power match intuitions while the resulting optimal graphs using the conjunctive power coincide with many existing multiple comparison procedures. The optimization process is efficient because it uses numerical optimization instead of simulations or more computationally intensive methods, e.g., deep learning.

One future direction to optimize graphical multiple comparison procedures is to tailor the objective function. Although the disjunctive and the conjunctive power are often used, they may not reflect clinical considerations and structures within a graph, e.g., secondary hypotheses are considered only if the corresponding primary hypothesis has been rejected. So a more tailored objective function could better align with the objective of a trial. One challenge in this direction is that it is not easy to find a single objective function that optimization over it could uniquely identify the optimal graph (e.g., neither the disjunctive nor the conjunctive power could). Further investigation is needed to better understand the behavior of objective functions and their link to hypothesis and transition weights.

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

References

1. European Medicines Agency. Guideline on multiplicity issues in clinical trials (Draft). Available at https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf; 2017. Accessed May 24, 2023.

2. Food and Drug Administration. Multiple endpoints in clinical trials Guidance for industry. Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry; 2022. Accessed May 24, 2023.

3. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine.* 2009;28(4):586–604.

4. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine.* 2009;28:739–761.

5. Maurer W, Bretz F. Memory and other properties of multiple test procedures generated by entangled graphs. *Statistics in Medicine.* 2013;32(10):1739–1753.

6. Millen BA, Dmitrienko A. Chain procedures: A class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research.* 2011;3(1):14–30.

7. Maurer W, Hothorn LA, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie.* 1995;6:3–18.

8. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics.* 1979:65–70.

9. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics.* 2003;2:211–215.

10. Wiens BL, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics.* 2005;15(6):929–942.

11. Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine.* 2003;22:2387–2400.

12. Dmitrienko A, Tamhane AC, Wiens BL. General multistage gatekeeping procedures. *Biometrical Journal.* 2008;50:667–677.

13. Xi D, Tamhane AC. A general multistage procedure for k-out-of-n gatekeeping. *Statistics in Medicine.* 2014;33(8):1321–1335.

14. Julious SA. Sample sizes for clinical trials with normal data. *Statistics in Medicine.* 2004;23(12):1921–1986.

15. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics.* 2007;6(3):161–170.

16. Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics.* 1997;24(3):407–418.

17. Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference.* 2001;99:25-40.

18. Spjøtvoll E. On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics.* 1972;43(2):398–411.

19. Zhang Z, Wang C, Troendle JF. Optimizing the order of hypotheses in serial testing of multiple endpoints in clinical trials. *Statistics in Medicine.* 2015;34(9):1467–1482.

20. Rubin D, Dudoit S, Laan V. dM. A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology.* 2006;5(1).

21. Roeder K, Wasserman L. Genome-wide significance levels and weighted hypothesis testing. *Statistical science.* 2009;24(4):398.

22. Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine.* 2011;30(13):1489–1501.

23. Wiens BL, Dmitrienko A, Marchenko O. Selection of hypothesis weights and ordering when testing multiple hypotheses in clinical trials. *Journal of Biopharmaceutical Statistics.* 2013;23(6):1403–1419.

24. Dmitrienko A, Paux G, Brechenmacher T. Power calculations in clinical trials with complex clinical objectives. *Journal of the Japanese Society of Computational Statistics.* 2015;28(1):15–50.

25. Zhan T, Hartford A, Kang J, Offen W. Optimizing graphical procedures for multiplicity control in a confirmatory clinical trial via deep learning. *Statistics in Biopharmaceutical Research.* 2022;14(1):921-102.

26. Chen J, Luo J, Liu K, Mehrotra DV. On power and sample size computation for multiple testing procedures. *Computational statistics & data analysis.* 2011;55(1):110–122.

27. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63(3):655–660.

28. Westfall PH, Krishen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. *Statistics in medicine.* 1998;17(18):2107–2119.

29. Kraft D. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, Institut für Dynamik der Flugsysteme, Oberpfaffenhofen, Germany; 1988.

30. Ypma J, et al. nloptr: R Interface to NLopt. https://CRAN.R-project.org/package=nloptr; 2022. R package version 2.0-3.

31. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association.* 1955;50(272):1096–1121.

32. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics.* 1949:99–114.

33. Genz A, et al. mvtnorm: Multivariate Normal and t Distributions. https://CRAN.R-project.org/package=mvtnorm; 2021. R package version 1.1-3.

34. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika.* 1986;73(3):751-754.

35. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika.* 1988;75(2):383–386.

36. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika.* 1988;75(4):800–802.

37. Gou J, Tamhane AC, Xi D, Rom D. A class of improved hybrid Hochberg–Hommel type step-up multiple test procedures. *Biometrika.* 2014;101(4):899–911.

38. Xi D, Bretz F. Symmetric graphs for equally weighted tests, with application to the Hochberg procedure. *Statistics in Medicine.* 2019;38(27):5268-5282.

39. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal.* 2011;53(6):894–913.

40. Sarkar SK. Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Ann Stat.* 1998:494-504.

41. Huque MF. Validity of the Hochberg procedure revisited for clinical trial applications. *Statist Med.* 2016;35(1):5-20.

42. Gou J, Tamhane AC. Hochberg procedure under negative dependence. *Stat Sin.* 2018;28:339-362.

43. Miwa T, Hayter AJ, Kuriki S. The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2003;65(1):223–234.

44. Genz A, Bretz F. Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics.* 2002;11(4):950–971.

APPENDIX

A PROOFS

A.1 Proofs for Section 3.1

Theorem 1. For null hypotheses $\{H_i\}$ with the marginal power $\{d_i\}$, $i = 1, \dots, m$ and independent test statistics, there exists a unique set of Bonferroni weights $\{w_i\}$ that optimizes the disjunctive power in (7), if

$$1 - \Phi\left[\Phi^{-1}(1-d) + \frac{\phi\{\Phi^{-1}(1-d)\}}{1-d}\right] > \alpha, \tag{A1}$$

where $d = \min\left\{d_i\right\}$ is the smallest marginal power.

Proof of Theorem 1. Because the disjunctive power is a continuous and differentiable function in the support of $w_i \in [0,1]$ for all $i = 1\ldots, m$, the optimal disjunctive power exists. Additionally, if $w_i = 0$ for any i, we have $P_{\varepsilon_i}\left(p_i \geq w_i\alpha\right) = 0$ and the disjunctive power is 0. Thus the optima is within the interior of the support $w_i \in (0,1)$. Let the disjunctive power in (7) be g. Thus, the optima point can be found by solving $\frac{dg}{dw_i} = 0$. For any i, we have

$$\frac{dg}{dw_i} = \frac{\Phi\left(\eta_i - C_i\right)\cdot\phi\left(\eta_i\right)}{\phi\left(\eta_i - C_i\right)},$$

where $\eta_i = \Phi^{-1}(1-w_i\alpha)$ and $C_i = \Phi^{-1}(1-\alpha) - \Phi^{-1}(1-d_i)$. To show the uniqueness of the optima, we need to show that $\frac{dg}{dw_i}$ is strictly monotone increasing in w_i. Because if it is, then $w_i > w_j$ leads to $\frac{dg}{dw_i} > \frac{dg}{dw_j}$. Thus, it is not possible to have another optima of $w_i' > w_i$ and $w_j' < w_j$ such that $\frac{dg}{dw_i'} > \frac{dg}{dw_i}$, $\frac{dg}{dw_j'} > \frac{dg}{dw_j}$ but $\frac{dg}{dw_i} = \frac{dg}{dw_j} = 0$.

To prove $\frac{dg}{dw_i}$ is strictly monotone increasing, we only need to show that $\frac{dg}{dw_i dw_i} > 0$ for $w_i, w_j \in (0,1)$. Because $\frac{d\eta_i}{dw_i} = \frac{-\alpha}{\phi(\eta_i)}$ and $\frac{d\phi(x)}{dx} = -x\cdot\phi(x)$, we have

$$\frac{dg}{dw_i dw_i} = \frac{-\alpha\cdot\phi^2(\eta_i - C_i) + \alpha\cdot C_i\cdot\Phi(\eta_i - C_i)\cdot\phi(\eta_i - C_i)}{\phi^2(\eta_i - C_i)}.$$

Since the denominator is always greater than 0, we only need to find the condition such that the nominator is greater then 0, i.e. when $C_i > \frac{\phi(\eta_i - C_i)}{\Phi(\eta_i - C_i)}$. Recall that $\eta_i = \Phi^{-1}(1-w_i\alpha)$ and $C_i = \Phi^{-1}(1-\alpha) - \Phi^{-1}(1-d_i)$, we have $\eta_i - C_i > \Phi^{-1}(1-d_i)$ for $w_i \in (0,1)$. Because $\frac{\phi(x)}{\Phi(x)}$ is decreasing, we have

$$C_i > \frac{\phi\left\{\Phi^{-1}(1-d_i)\right\}}{1-d_i} > \frac{\phi\left\{\Phi^{-1}(1-d)\right\}}{1-d} \text{ for } i = 1,\ldots,m,$$

where $d = \min\left\{d_i\right\}$. The above condition can be further strengthened as

$$\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-d) > \frac{\phi\left\{\Phi^{-1}(1-d)\right\}}{1-d}.$$

Therefore, a sufficient condition for $\frac{dg}{dw_i}$ being strictly monotone increasing and thus for the uniqueness of the optima for (7) is

$$1 - \Phi\left[\Phi^{-1}(1-d) + \frac{\phi\{\Phi^{-1}(1-d)\}}{1-d}\right] > \alpha.$$

□

A.2 Proofs for Section 4.1

Theorem 2. For null hypotheses $\{H_i\}$ with the marginal power $\{d_i\}$, $i = 1,\ldots,m$ and independent test statistics, there exists a unique set of Bonferroni weights $\{w_i\}$ that optimizes the conjunctive power in (10), if

$$1 - \Phi\left[\Phi^{-1}(1-d) + \frac{\phi\{\Phi^{-1}(1-d)\}}{1-d}\right] > \alpha, \tag{A2}$$

where $d = \min\left\{d_i\right\}$ is the smallest marginal power.

Proof of Theorem 2. Because the conjunctive power is a continuous and differentiable function in the support of $w_i \in [0,1]$ for all $i = 1\ldots, m$, the optimal conjunctive power exists. Additionally, if $w_i = 0$ for any i, we have $P_{\varepsilon_i}\left(p_i \geq w_i\alpha\right) = 0$ and the conjunctive power is 0. Thus the optima is within the interior of the support $w_i \in (0,1)$. Let the conjunctive power in (10) be h. Thus, the optima point can be found by solving $\frac{dh}{dw_i} = 0$. For any i, we have

$$\frac{dh}{dw_i} = \frac{1 - \Phi\left(\eta_i - C_i\right)\cdot\phi\left(\eta_i\right)}{\phi\left(\eta_i - C_i\right)},$$

where $\eta_i = \Phi^{-1}(1-w_i\alpha)$ and $C_i = \Phi^{-1}(1-\alpha) - \Phi^{-1}(1-d_i)$.

From the proof of Theorem 1, a sufficient condition for $\frac{dh}{dw_i}$ being strictly monotone decreasing and thus for the uniqueness of the optima for (10) is

$$1 - \Phi\left[\Phi^{-1}(1-d) + \frac{\phi\{\Phi^{-1}(1-d)\}}{1-d}\right] > \alpha.$$

□

B COMPUTATIONAL CONSIDERATIONS

There is a trade-off between precision and computing time. The first computational consideration is about the numerical calculation of probabilities under multivariate normal distributions. Two algorithms have been considered in the `mvtnorm` R package. The Miwa algorithm[43] is a deterministic algorithm which is fast when the number of hypotheses is up to 20. However, it does not handle singular cases and may not be stable for all correlation structures. The GenzBretz algorithm[44] is a quasi randomized Monte-Carlo procedure, which is applicable to arbitrary correlation structures and dimensions up to 1000. We adopt the GenzBretz algorithm with the default setting of precision because a higher precision requires much more computing time.

The second computational consideration is about the choice of numerical optimization algorithms. We consider the `nloptr` R package which provides a library of non-linear optimization algorithms. Gradient-based algorithms utilize gradients of the objective function and thus are more computationally efficient than non-gradient-based algorithms. Thus the former requires additional information of gradients while the latter does not. In addition, our optimization problems requires an equality constraint and box constraints. Given these considerations, we narrow down our candidate algorithms to the Sequential Least SQuares Programming (SLSQP) algorithm, the Improved Stochastic Ranking Evolution Strategy (ISRES), and the Constrained Optimization BY Linear Approximation (COBYLA) algorithm. The comparison among these three algorithms is provided in Table B1. Since SLSQP is easy to implement with gradients and more importantly the fastest algorithm and thus it is chosen for numerical optimization.

Algorithm	Gradient based	Computing time	Equality constraint	Multiple starting points needed
SLSQP	Yes	Short	Yes	Yes
ISRES	Yes	Long	Yes	Yes
COBYLA	No	Medium	Needs transformation	Yes

TABLE B1 Comparison of candidate optimization algorithms.

Finally, we evaluate computing time as a function of the number of hypotheses for the optimal disjunctive power using Algorithm (1). The comparison is conducted using Intel(R) Xeon(R) Platinum CPU @ 3.00GHz. Three scenarios are considered including (1) the independent case with random marginal power values drawn from a uniform distribution from 0.5 to 0.99, (2) the dependent case with an equal correlation drawn from uniform distribution from 0 to 1 and a single starting point of equal weights, and (3) the dependent case with multiple starting points in Algorithm (1). The computing time based on 100 simulations are presented in Table B2 for 2 to 10 hypotheses. Note that the computing time for Algorithm (1) with multiple starting points has more variability, depending on the marginal power and the correlation structure. For example the cases with 10 hypotheses, the computing time ranges from less than 5 minutes to more than one hour due to possible early stopping in the optimization algorithm.

C BONFERRONI WEIGHTS WITH TWO HYPOTHESES

Consider two hypotheses H_1, H_2 and apply a weighted Bonferroni test to control the FWER at one-sided level $\alpha = 0.025$. For the marginal power, we consider both equal and unequal scenarios. Assuming independence between test statistics, optimal weights that maximize the disjunctive power are provided in Table C3. Same conclusions can be made as those with three hypotheses in Section 3. When the marginal power is equal for H_1, H_2, the optimal weights are always 0.5. When the marginal power is

Number of hypotheses	Scenario (1)	(2)	(3)	Number of starting points for (3)
2	0.0025	0.0074	0.0229	3
4	0.0048	0.3726	3.6467	15
6	0.0107	1.7702	55.9015	63
8	0.0151	3.9565	398.6857 (~6.6 mins)	255
10	0.0238	6.4554	3291.107 (~55 mins)	1023

TABLE B2 Computing time in seconds of the proposed algorithms.

different for different hypotheses, the one with the higher marginal power receives a larger optimal weight and the one with the lower marginal power receives a smaller optimal weight. In all cases, the optimal disjunctive power is at least as large as the largest marginal power.

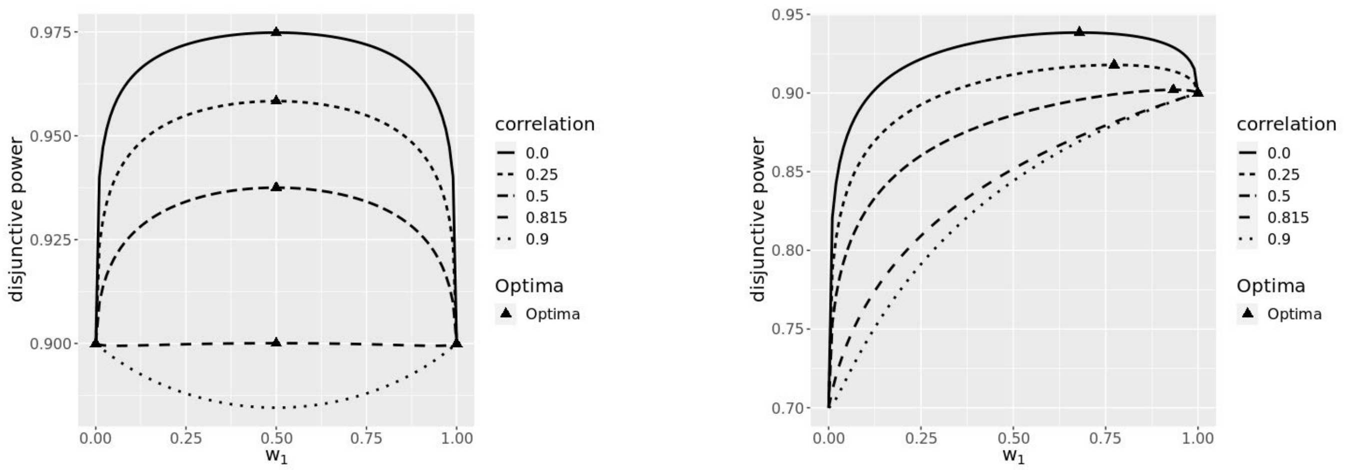
Marginal power (%) H_1, H_2	Correlation ρ	Weights w_1	w_2	Disjunctive power (%)
90, 90	0	0.5	0.5	97.484
80, 80	0	0.5	0.5	91.724
90, 80	0	0.597	0.403	95.495
90, 70	0	0.679	0.321	93.842
90, 50	0	0.829	0.171	91.424

TABLE C3 Optimal Bonferroni weights for the disjunctive power with two hypotheses under independence.

To assess the impact of the correlation on the optimal weights, we consider two sets of the marginal power: (90%, 90%) and (90%, 70%). In Figure C1, we plot the disjunctive power against the Bonferroni weight w_1 for five values of the correlation from 0 to 0.9. In panel (a) of Figure C1, we consider the marginal power of (90%, 90%). As the correlation increases from 0 to 0.9, the disjunctive power curve is always symmetric around $w_1 = 0.5$; its value decreases from above 90% to below 90%; and its shape changes from being concave to being convex. In particular, when the correlation is about 0.815, the disjunctive power is almost constant at 90%. Therefore, when the correlation is lower than 0.815, the optimal Bonferroni weight is $w_1 = 0.5$; when the correlation is higher than 0.815, the optimal Bonferroni weight is $w_1 = 1$ or $w_1 = 0$. In panel (b) of Figure C1, we consider the marginal power of (90%, 70%). As the correlation increases from 0 to 0.9, the disjunctive power decreases. The optima is denoted as the node on each curve. We can see that $w_1 > 0.5$ in all cases; it increases in the correlation; it is 1 when the correlation is 0.9.

Next consider the conjunctive power for two hypotheses. For the marginal power, we consider both equal and unequal scenarios. Assuming independence between test statistics, optimal weights that maximize the conjunctive power are provided in Table C4. Same conclusions can be made as those with three hypotheses in Section 4. When the marginal power is equal for H_1, H_2, the optimal weights are always 0.5, regardless of the exact number of the marginal power. When the marginal power is different for different hypotheses, the one with the higher marginal power receives a smaller optimal weight and the one with the lower marginal power receives a larger optimal weight. In all cases, the optimal conjunctive power is less than the smallest marginal power.

To assess the impact of the correlation on the optimal weights, we consider two sets of the marginal power: (90%, 90%) and (90%, 70%). In Figure C2, we plot the conjunctive power against the Bonferroni weight w_1 for five values of the correlation from 0 to 0.9. In panel (a) of Figure C2, we consider the marginal power of (90%, 90%). As the correlation increases from 0 to 0.9, the conjunctive power curve is always symmetric around $w_1 = 0.5$; its value increases in the correlation; the optimal Bonferroni weight is always $w_1 = 0.5$. In panel (b) of Figure C2, we consider the marginal power of (90%, 70%). As the correlation increases from 0 to 0.9, the conjunctive power increases. The optima is denoted as the node on each curve. We can see that $w_1 < 0.5$ in all case.



(a) Equal marginal power of 90% **(b)** Unequal marginal power of 90% and 70%

FIGURE C1 Impact of correlation on the disjunctive power of two hypotheses under different w_1.

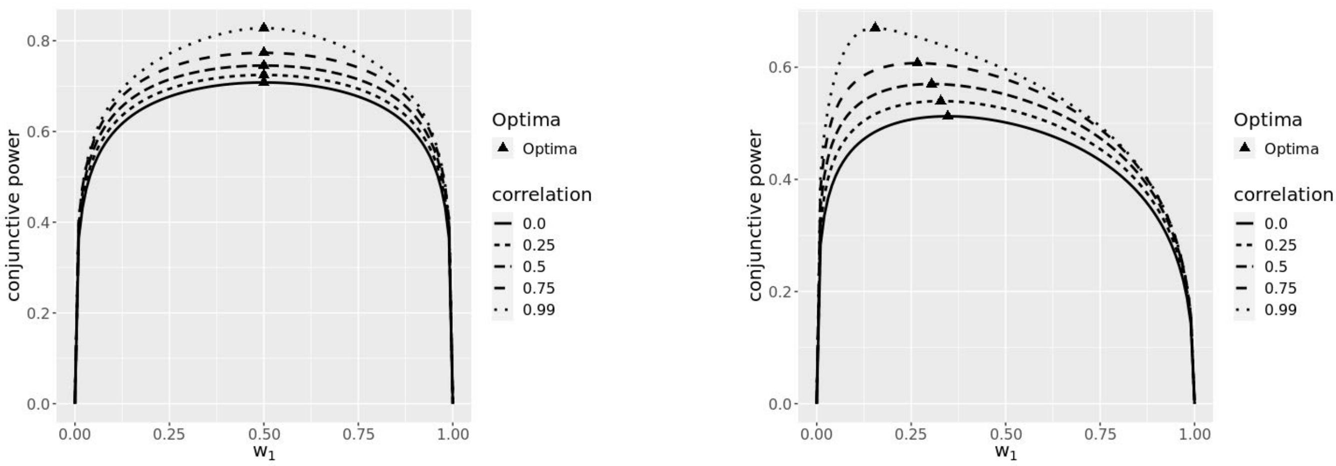
Marginal power (%)	Correlation	Weights		Conjunctive
H_1, H_2	ρ	w_1	w_2	power (%)
90, 90	0	0.5	0.5	70.791
80, 80	0	0.5	0.5	50.740
90, 80	0	0.404	0.596	60.361
90, 70	0	0.347	0.653	51.252
90, 50	0	0.273	0.727	34.912

TABLE C4 Optimal Bonferroni weights for the conjunctive power with two hypotheses under independence.

D BONFERRONI WEIGHTS WITH FOUR HYPOTHESES

Consider four hypotheses H_1, H_2, H_3, H_4 and apply a weighted Bonferroni test to control the FWER at one-sided level $\alpha = 0.025$. For the marginal power, we consider both equal and unequal scenarios. Assuming independence between test statistics, optimal weights that maximize the disjunctive power are provided in Table D5. Same conclusions can be made as those with three hypotheses in Section 3. When the marginal power is equal for H_1, H_2, H_3, H_4, the optimal weights are always 0.25. In addition, the optimal weights are also equal as long as the marginal power is constant for a subset of hypotheses, e.g., for only H_2, H_3, H_4 in the last two scenarios. When the marginal power is different for different hypotheses, the one with the higher marginal power receives a larger optimal weight and the one with the lower marginal power receives a smaller optimal weight. In all cases, the optimal disjunctive power is at least as large as the largest marginal power.

When test statistics are correlated, the marginal power and the correlation matrix could have a joint effect on optimal weights. For four hypotheses H_1, H_2, H_3, H_4, we consider constant and non-constant marginal power values. For the correlation matrix, both homogeneous and heterogeneous structures are assessed. Heterogeneous structures are provided in matrices (D3) and (D4). Optimal weights that maximize the disjunctive power are provided in Table D6. For homogeneous correlation structure, we observe a similar trend as the independence case in Table D5: the higher the marginal power, the larger the optimal weight. As the correlation increases, this unequal split becomes more prevail and more extreme. For example, when the marginal power is (90%, 75%, 60%. 45%) for H_1, H_2, H_3, H_4 respectively, the optimal weight is (0.502, 0.279, 0.152, 0.066) under independence, (0.866, 0.134, 0, 0) under correlation of 0.5, and (1, 0, 0, 0) under correlation above 0.7. When the correlation increase between two hypotheses, the optimal weight increases for the hypothesis with a larger marginal power and decrease for the one with a lower marginal power. For example, when the marginal power is (90%, 75%, 60%, 45%) for H_1, H_2, H_3, H_4 respectively, the optimal weight for H_1 increases from 0.866 to 1 when ρ_{12} changes from 0.5 to 0.8 and the optimal weight for H_2 decreases from 0.134 to 0.000. The gain in the disjunctive power from optimal weights under dependence ranges from 0.2% to 5.9%, compared



(a) Equal marginal power of 90%

(b) Unequal marginal power of 90% and 70%

FIGURE C2 Impact of correlation on the conjunctive power of two hypotheses under different w_1.

Marginal power (%)	Correlation	Weights				Disjunctive
H_1, H_2, H_3, H_4	ρ	w_1	w_2	w_3	w_4	power (%)
90, 90, 90, 90	0	0.25	0.25	0.25	0.25	99.727
80, 80, 80, 80	0	0.25	0.25	0.25	0.25	97.901
90, 80, 70, 60	0	0.415	0.276	0.187	0.122	97.519
90, 75, 60, 45	0	0.502	0.279	0.152	0.066	96.024
90, 70, 50, 10	0	0.677	0.320	0.003	0	93.847
90, 50, 50, 50	0	0.633	0.122	0.122	0.122	93.417
90, 10, 10, 10	0	1	0	0	0	90.000

TABLE D5 Optimal Bonferroni weights for the disjunctive power with four hypotheses under independence.

to that from optimal weights under independence (benchmarks). For heterogeneous correlation structures, the trend of optimal weights comparing with benchmarks depends on the detailed correlation structure. For example, in the last two rows of Table D6, optimal weight of w_2 is smaller than w_3 even the marginal power of H_2 is higher than H_3. The reason is the correlation between H_1 and H_2 is much higher than the one between H_1 and H_3, which causes $w_2 = 0$ when H_1 presents. However, H_3 still get the non-zero optimal weight when H_1 exists, because of its approximate independence with H_1.

$$\Sigma_1 = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 \\ 0.8 & 1 & 0.6 & 0.4 \\ 0.6 & 0.6 & 1 & 0.4 \\ 0.4 & 0.4 & 0.4 & 1 \end{bmatrix} \tag{D3}$$

$$\Sigma_2 = \begin{bmatrix} 1 & 0.9 & 0.1 & 0.4 \\ 0.9 & 1 & 0.1 & 0.4 \\ 0.1 & 0.1 & 1 & 0.4 \\ 0.4 & 0.4 & 0.4 & 1 \end{bmatrix} \tag{D4}$$

Next consider the conjunctive power for four hypotheses. For the marginal power, we consider both equal and unequal scenarios. Assuming independence between test statistics, optimal weights that maximize the conjunctive power are provided in Table D7. Same conclusions can be made as those with three hypotheses in Section 4. When the marginal power is equal for H_1, H_2, H_3, H_4, the optimal weights are always 0.25. In addition, the optimal weights are also equal as long as the marginal power is constant for a subset of hypotheses, e.g., for only H_2, H_3, H_4 in the last two scenarios. When the marginal power is different for different hypotheses, the one with the higher marginal power receives a smaller optimal weight and the one with

Marginal power (%) H_1, H_2, H_3, H_4	Correlation	Optimal weights w_1	w_2	w_3	w_4	Disjunctive power (%)
(90, 90, 90, 90)	0.9	1	0	0	0	90.000
	Benchmark	0.25	0.25	0.25	0.25	86.384
	0.8	0.5	0.5	0	0	90.238
	Benchmark	0.25	0.25	0.25	0.25	89.825
	0.77	1/3	1/3	1/3	0	90.725
	Benchmark	0.25	0.25	0.25	0.25	90.631
	0.7	0.25	0.25	0.25	0.25	92.273
	0.5	0.25	0.25	0.25	0.25	95.730
	Σ_1	0	0.288	0.288	0.424	95.499
		0.288	0	0.288	0.424	95.499
	Benchmark	0.25	0.25	0.25	0.25	95.272
	Σ_2	0.411	0	0.411	0.178	97.305
		0	0.411	0.411	0.178	97.305
	Benchmark	0.25	0.25	0.25	0.25	96.954
(80, 80, 80, 80)	0.9	1	0	0	0	80.000
	Benchmark	0.25	0.25	0.25	0.25	74.077
	0.8	1	0	0	0	80.000
	Benchmark	0.25	0.25	0.25	0.25	78.982
	0.77	0.5	0.5	0	0	80.584
	Benchmark	0.25	0.25	0.25	0.25	80.185
	0.7	0.25	0.25	0.25	0.25	82.701
	0.5	0.249	0.25	0.25	0.25	88.479
	Σ_1	0	0.287	0.287	0.426	88.087
		0.287	0	0.287	0.426	88.087
	Benchmark	0.25	0.25	0.25	0.25	87.559
	Σ_2	0.410	0	0.410	0.180	91.441
		0	0.410	0.410	0.180	91.441
	Benchmark	0.25	0.25	0.25	0.25	90.555
(90, 75, 60, 45)	0.9	1	0	0	0	90.000
	Benchmark	0.502	0.279	0.152	0.066	84.375
	0.8	1	0	0	0	90.000
	Benchmark	0.502	0.279	0.152	0.066	85.208
	0.7	1	0	0	0	90.000
	Benchmark	0.502	0.279	0.152	0.066	86.351
	0.5	0.866	0.134	0	0	90.538
	Benchmark	0.502	0.279	0.152	0.066	88.997
	Σ_1	0.999	0	0	0.001	90.001
	Benchmark	0.502	0.279	0.152	0.066	86.223
	Σ_2	0.796	0	0.204	0	91.745
	Benchmark	0.502	0.279	0.152	0.066	88.793

TABLE D6 Optimal Bonferroni weights for the disjunctive power with four hypotheses under dependence.

the lower marginal power receives a larger optimal weight. In all cases, the optimal conjunctive power is less than the smallest marginal power.

When test statistics are correlated, we consider constant and non-constant marginal power values. For the correlation matrix, both homogeneous and heterogeneous structures are assessed. Heterogeneous structures are provided in matrices (D3) and (D4). Optimal weights that maximize the conjunctive power are provided in Table D8. For homogeneous correlation structure, we observe a similar trend as the independence case in Table D7: the higher the marginal power, the smaller the optimal weight.

Marginal power (%)	Correlation	Weights				Conjunctive
H_1, H_2, H_3, H_4	ρ	w_1	w_2	w_3	w_4	power (%)
90, 90, 90, 90	0	0.25	0.25	0.25	0.25	35.429
80, 80, 80, 80	0	0.25	0.25	0.25	0.25	14.718
90, 80, 70, 60	0	0.158	0.224	0.281	0.337	9.579
90, 75, 60, 45	0	0.136	0.217	0.287	0.360	4.558
90, 70, 50, 10	0	0.104	0.181	0.250	0.465	0.499
90, 50, 50, 50	0	0.121	0.293	0.293	0.293	2.163
90, 10, 10, 10	0	0.074	0.309	0.309	0.309	0.004

TABLE D7 Optimal Bonferroni weights for the conjunctive power with four hypotheses under independence.

As the correlation increases, this unequal split becomes more prevail and more extreme. For example, when the marginal power is (90%, 75%, 60%. 45%) for H_1, H_2, H_3, H_4 respectively, the optimal weight under independence is is (0.136, 0.217, 0.287, 0.360), (0.087, 0.186, 0.295, 0.432) under correlation of 0.5, and (0.039, 0.135, 0.289, 0.536) under correlation 0.9. When the correlation increase between two hypotheses, the optimal weight decreases for the hypothesis with a larger marginal power and increases for the one with a lower marginal power. For example, when the marginal power is (90%, 75%, 60%, 45%) for H_1, H_2, H_3, H_4 respectively, the optimal weight for H_1 decreases from 0.087 to 0.055 when ρ_{12} changes from 0.5 to 0.8 and the optimal weight for H_4 increases from 0.432 to 0.497. The gain in the conjunctive power from optimal weights under dependence ranges from 0.2% to 2.4%, compared to that from optimal weights under independence (benchmarks).

E RATIONALES FOR STEP 5 OF ALGORITHM 2

In Algorithm 2, it is undefined to calculate the transition weight $g_{ji} = \left(w_i' - w_i\right)/w_i$ when $w_i = 0$. A practical solution is to set $w_i = \varepsilon > 0$, which is a small value and then normalize the rest of hypothesis weights. Alternative attempt has been tried to reject hypotheses until H_i has a positive weight and then to calculate its transition weights. However, we show in this example that the rejection order will change the transition weights from the optimization algorithm. A graph with four hypotheses is illustrated in Figure E3. Initially, H_3 and H_4 have 0 weight and all 0 transition weight have been removed from the graph. Let the marginal power be 90%, 70%, 80%, and 80%, respectively for H_1, H_2, H_3, H_4.

In the first step, the optimal Bonferroni weights for H_1 and H_2 are 0.679 and 0. 321, respectively to maximize the disjunctive power. In the second step, the transitions weights going out of H_1 and H_2 are easy to calculated based on Steps 6 and 7 in Algorithm 2, as shown in Figure E3. However the transition weights going out of H_3 and H_4, i.e. g_{31}, g_{32} and g_{34}, are not defined since $w_3 = w_4 = 0$.

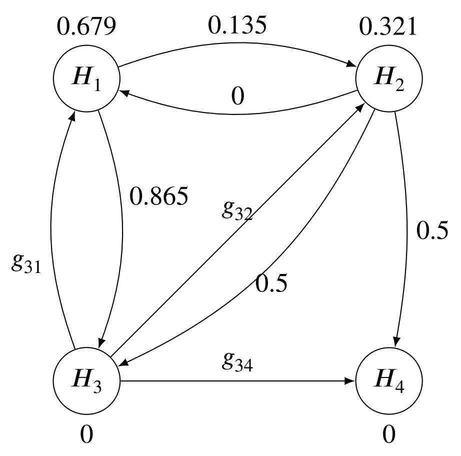


FIGURE E3 Example graph to illustrate rationales for Step 5 of Algorithm 2.

Marginal power (%) H_1, H_2, H_3, H_4	Correlation	Optimal weights w_1	w_2	w_3	w_4	Conjunctive power (%)
(90, 90, 90, 90)	0.9	0.25	0.25	0.25	0.25	66.633
	0.8	0.25	0.25	0.25	0.25	61.813
	0.77	0.25	0.25	0.25	0.25	60.570
	0.7	0.25	0.25	0.25	0.25	57.875
	0.5	0.25	0.25	0.25	0.25	51.068
	Σ_1	0.224	0.224	0.255	0.297	52.800
	Benchmark	0.25	0.25	0.25	0.25	52.628
	Σ_2	0.217	0.217	0.308	0.257	49.773
	Benchmark	0.25	0.25	0.25	0.25	49.520
(80, 80, 80, 80)	0.9	0.25	0.25	0.25	0.25	49.135
	0.8	0.25	0.25	0.25	0.25	43.503
	0.77	0.25	0.25	0.25	0.25	42.070
	0.7	0.25	0.25	0.25	0.25	38.990
	0.5	0.249	0.25	0.25	0.25	31.367
	Σ_1	0.220	0.220	0.255	0.304	33.083
	Benchmark	0.25	0.25	0.25	0.25	32.885
	Σ_2	0.215	0.215	0.317	0.253	29.292
	Benchmark	0.25	0.25	0.25	0.25	29.031
(90, 75, 60, 45)	0.9	0.039	0.135	0.289	0.536	29.291
	Benchmark	0.136	0.217	0.287	0.360	26.878
	0.8	0.055	0.155	0.294	0.497	24.936
	Benchmark	0.136	0.217	0.287	0.360	23.630
	0.7	0.067	0.168	0.296	0.470	21.447
	Benchmark	0.136	0.217	0.287	0.360	20.676
	0.5	0.087	0.186	0.295	0.432	15.675
	Benchmark	0.136	0.217	0.287	0.360	15.400
	Σ_1	0.065	0.169	0.297	0.469	15.365
	Benchmark	0.136	0.217	0.287	0.360	14.766
	Σ_2	0.068	0.198	0.336	0.398	12.871
	Benchmark	0.136	0.217	0.287	0.360	12.496

TABLE D8 Optimal Bonferroni weights for the conjunctive power with four hypotheses under dependence.

One thought would be to calculate the transition weights until the hypothesis receives a positive weight. In this case, H_3 will get a positive weight whenever one of H_1 and H_2 is rejected. When H_1 is rejected first, its weight of 0.679 will be propagated to H_2 and H_3 with the proportion of 0.135 and 0.865, respectively. Thus H_3 gets a weight 0.587. Then using Algorithm 2 to calculate transition weights going out of H_3, the optimal transition weights are $g'_{32} = \frac{g_{32}+g_{31}g_{12}}{1-g_{31}g_{13}} = 0$ and $g'_{34} = \frac{g_{34}+g_{31}g_{14}}{1-g_{31}g_{13}} = 1$, where $'$ indicates the updated graph after H_1 is removed. Thus $g_{32} = g_{31} = 0$ and $g_{34} = 1$.

However, if H_2 is rejected first, its weight of 0.321 will be propagated to H_3 and H_4 with the proportion of 0.5 and 0.5, respectively. Thus H_3 gets a weight 0.1605. Then using Algorithm 2 to calculate transition weights going out of H_3, the optimal transition weights are $g'_{31} = \frac{g_{32}+g_{32}g_{21}}{1-g_{32}g_{23}} = \frac{g_{32}}{1-g_{32}/2} = 0.474$ and $g'_{34} = \frac{g_{34}+g_{32}/2}{1-g_{32}/2} = 0.526$. Thus $g_{32} = 0.383$ and $g_{34} = 0.234$. However, these do not align with the above calculation. Therefore, we adopted a practical solution to in Step 5 of Algorithm 2 to get around with the zero weight.