Admissibility in Partial Conjunction Testing

Jingshu Wang [*]

Department of Statistics, Univeristy of Pennsylvania
and
Art B. Owen [†]
Department of Statistics, Stanford University

May, 2017

Abstract

Meta-analysis combines results from multiple studies aiming to increase power in finding their common effect. It would typically reject the null hypothesis of no effect if any one of the studies shows strong significance. The partial conjunction null hypothesis is rejected only when at least r of n component hypotheses are non-null with $r = 1$ corresponding to a usual meta-analysis. Compared with meta-analysis, it can encourage replicable findings across studies. A by-product of it when applied to different r values is a confidence interval of r quantifying the proportion of non-null studies. Benjamini and Heller (2008) provided a valid test for the partial conjunction null by ignoring the $r-1$ smallest p-values and applying a valid meta-analysis p-value to the remaining $n-r+1$ p-values. We provide sufficient and necessary conditions of admissible combined p-value for the partial conjunction hypothesis among monotone tests. Non-monotone tests always dominate monotone tests but are usually too unreasonable to be used in practice. Based on these findings, we propose a generalized form of Benjamini and Heller's test which allows usage of various types of meta-analysis p-values, and apply our method to an example in assessing replicable benefit of new anticoagulants across subgroups of patients for stroke prevention.

Keywords: meta-analysis, replicable findings, combined p-values, power, subgroup analysis

[*] jingshuw@wharton.upenn.edu, supported in part by NIH grant 2 R01 HG006137, and the Wharton Dean's Fund

[†] owen@stanford.edu, supported in part by NSF grants DMS-1407397 and DMS-1521145

Table 1: four hypothetical cases for five ordered p-values.

Case	$p_{(1)}$	$p_{(2)}$	$p_{(3)}$	$p_{(4)}$	$p_{(5)}$
A	10^{-200}	0.4	0.5	0.6	0.7
B	10^{-10}	10^{-9}	10^{-8}	10^{-7}	10^{-6}
C	10^{-100}	10^{-100}	10^{-100}	0.049	0.8
D	0.048	0.048	0.048	0.048	0.8

1 Introduction

When a null hypothesis is tested in n different settings, a meta-analysis can be used to obtain a combined p-value based on all of the test results. It gains power as the combined p-value is usually more significant than any of the individual p-value in each setting. However, the combined p-value in meta-analysis is only valid for the global null where the null hypothesis is true in every setting, thus it is possible that the null is then rejected largely on the basis of just one extremely significant component hypothesis test. Such a rejection may be undesirable as it could arise from some irreproducible property of the setting in which that one component test was made.

Refering to Table 1, cases A and B illustrate a potential problem for meta-analysis. Both a Fisher and a Stouffer meta-analysis would find case A more significant than case B, although the only significant setting in case A may be largely due to a technical or statistical bias. The random effect model in meta-analysis has been widely accepted for consideration of heterogeneity across studies (Higgins et al., 2009). However, it still assumes that the effects across studies are similar and does not explicitly guarantee replication nor robustness to extreme bias.

Researchers in functional magnetic resonance imaging (fMRI) have adopted conjunction (logical 'and') testing (Price and Friston, 1997; Friston et al., 1999; Nichols et al., 2005) in which a hypothesis must be rejected in all n settings where it is tested. The n settings may correspond to related tasks or they may correspond to independent subjects. For example in Table 1, a conjunction test would only reject case B which shows consistent replication. However, conjunction tests lose power for large n as they are based on the largest of n p-values. A compromise is to require evidence that at least r out of n null hypotheses are false, for some user specified r. Such tests of the 'partial conjunction (PC) null hypothesis' were used in Friston et al. (2005) and then studied by Benjamini and Heller (2008). The extremes $r = 1$ and $r = n$ correspond to the usual meta-analysis tests and conjunction testing respectively.

Partial conjunction testing is useful in areas beyond neuroimaging. It has been applied in systematic reviews of preventitive healthcare (Shenhav et al., 2015) and genome-wide association studies (Heller and Yekutieli, 2014). The PC test has potential usage in finding common gene regulation patterns across tissues for eQTL data (Flutre et al., 2013). Finally, it has been applied to gene set enrichment analysis to avoid selection of gene sets whose significance depend on only one single gene (Wang et al., 2010).

A Benjamini-Heller partial conjunction (BHPC) test works as follows. One sorts the observed p-values yielding $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$, ignores the smallest $r-1$ of them, and then applies a valid p-value combination rule to the remaining $n-r+1$ p-values. Benjamini and Heller (2008) show that BHPC tests are valid for the partial conjunction null when the n hypotheses are independent. They also consider some dependent test conditions as well as the consequences of using PC tests in the Benjamini-Hochberg procedure.

Cases C and D illustrate an interesting property of the BHPC tests. Suppose that we need to reject at least four null hypotheses to have a meaningful finding. Then a BHPC test finds that case D is stronger evidence (smaller p-value) than case C, because BHPC is based only on $p_{(4)}$ and $p_{(5)}$. In case C we are extremely confident of three rejections and are banking on the fourth one to be correct. In case D by contrast, none of the four smallest p-values is much better than borderline. It appears to have about four times as many ways to disappoint us. This comparsion between case C and D reveals a counter-intuitive property of the BHPC tests, that we study at the end of this section.

In this paper, we investigate the power properties of BHPC tests focussing on admissibility. Under the assumption that the component p-values are either independent or have a positive dependence structure, we characterize the complete class of tests for monotone admissibility, which is a generalized form of BHPC p-values (GBHPC p-values). The only admissible PC tests among monotone tests are either of the BHPC form, or its generalization (GBHPC), which uses combined p-values constructed by taking the maximum of the meta-analysis p-value of each of the $\binom{n}{r-1}$ subsets of $n-r+1$ hypotheses. Under mild assumptions, a sufficient condition for the monotone admissibility of a GBHPC p-value is that each of the meta-analysis p-values for $\binom{n}{r-1}$ subsets is admissible. GBHPC p-values are also called r-values in Shenhav et al. (2015).

The monotonicity condition, which means that the combined p-value is a non-decreasing function of the individual p-values, is necessary for us to discuss admissibility for partial conjunction hypotheses with $r > 1$. If we relax this condition, then BHPC tests become inadmissible. Because non-monotone tests are quite unreasonable scientifically in most cases, this is not a strong criticism of BHPC. We side with Perlman and Wu (1999) in rejecting the admissibility criterion, and not the test, when methods lacking face-value validity are included in comparisons.

Given the admissibility properties of BHPC p-values, how do we explain their puzzling behavior for cases C and D in Table 1? An explanation is that unlike the combined p-values in meta-analysis, the PC p-values measure the strength of replicability (the true proportion of non-null studies) instead of the magnitude of effect size. A PC p-value can be much smaller when the true number (r_0) rather than the effect size of non-null studies is large. Compared with case D, the three extreme p-values of case C in Table 1 gain us much stronger evidence of a large effect size but not much more evidence on $r_0 \geq 4$. Thus, the PC p-values for both case C and case D are similar.

Section 2 presents our notation and some background on partial conjunction tests and admissibility. Section 3 proposes the GBHPC p-values and presents the main theorems on monotone admissible partial conjunction p-values. Section 4 uses simulations to compare the power of several GBHPC p-values under various hypothesis configurations. Compared with BHPC p-values, GBHPC p-values have the advantage that they can be constructed from more sophisticated meta-analysis p-values. We illustrate this benefit in Section 5 in an application of GBHPC p-values for assessing replicable benefit and safety concerns of new oral anticoagulants across subgroups of patients for stroke prevention. Section 6 has our conclusions and states some future work.

2 Preliminaries

2.1 Definitions and notations

The problem begins with n null hypotheses to test, H_{0i} for $i = 1, \dots, n$. Each H_{0i} is the hypothesis to test for an individual setting/study. The corresponding alternative hypotheses are H_{1i}. The i'th hypothesis refers to a parameter θ_i. If H_{0i} holds then $\theta_i \in \Theta_{0i}$, while H_{1i} specifies that $\theta_i \in \Theta_{1i}$. The parameter space for the i'th hypothesis is $\Theta_i = \Theta_{0i} \cup \Theta_{1i}$ and of course $\Theta_{0i} \cap \Theta_{1i} = \varnothing$. The parameter space of $(\theta_1, \dots, \theta_n)$ is $\Theta = \prod_i \Theta_i$.

To each hypothesis, there corresponds a p-value, p_i. There may be a loss of information in reducing a data set to one p-value. Yet often that loss is small and very commonly the researchers who gathered the original data share only their p-values for reasons that may include privacy of their subjects.

We use p_i to denote the numerical value of the p-value for the i'th hypothesis. It is the observed value of a corresponding random variable P_i. The sorted p-values are $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ and $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$ are the sorted random variables. Probability and expectation for functions of P_i are given by $\mathbb{P}_{\theta_i}$ and $\mathbb{E}_{\theta_i}$ respectively. We let $\theta = (\theta_1, \dots, \theta_n)$ and $\boldsymbol{P} = (P_1, \dots, P_n)$. Probability and expectation for functions of $\boldsymbol{P}$ are given by $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$. Each p_i is a valid P-value according to the definition below:

Definition 1 (Validity)**.** A valid component p-value P_i satisfies $\sup_{\theta_i \in \Theta_{0i}} \mathbb{P}_{\theta_i}(P_i \leq \alpha) \leq \alpha$ for all $0 \leq \alpha \leq 1$.

Besides being independent, positive dependence can be a common dependence structure across studies, especially when they share samples. For $P_1, \ldots, P_n$, we assume that they are positively associated (Esary et al., 1967) under any parameter $\theta \in \Theta$.

Definition 2 (Positively associated random variables)**.** Random variables $X_1, X_2, \cdots, X_n$ are positively associated if

$$\mathrm{Cov}(f(X_1, \cdots, X_n), g(X_1, \cdots, X_n)) \geq 0$$

for all bounded functions, $f(\cdot)$ and $g(\cdot)$, that are nondecreasing in each argument.

It is obvious that independent P-values are associated. We will use two properties of associated P-values from Esary et al. (1967). First, for any $(p_1, \cdots, p_n) \in [0, 1]^n$,

$$\mathbb{P}_\theta(P_1 \leq p_1, \cdots, P_n \leq p_n) \geq \prod_i \mathbb{P}_{\theta_i}(P_i \leq p_i). \tag{1}$$

For the second property, we say that a set $D \subset \mathbb{R}^n$ is nonincreasing, if $X \in D$ and $\tilde{X} \leq X$ (componentwise) implies that $\tilde{X} \in D$ too. Then for any nonincreasing sets D_1 and D_2, associated P-values satisfy

$$\mathbb{P}_\theta(\boldsymbol{P} \in D_1 \mid \boldsymbol{P} \in D_2) \geq \mathbb{P}_\theta(\boldsymbol{P} \in D_1). \tag{2}$$

Here we provide two examples where the P-values can be positively associated.

Example 1 (Multivariate normal test statistics)**.** Assume that the test statistics of the n studies follows a multivariate Gaussian distribution: $(T_1, T_2, \cdots, T_n) \sim \mathcal{N}(\mu, \Sigma)$ and $P_i = \mathbb{P}(T_i \geq T_i^{\text{obs}}; \mu)$ is a one-sided P-value for μ_i. The one-sided P-values are either all left-sided or right-sided. Then the P_i are positively associated if for each $i \neq j$, $\Omega_{ij} \leq 0$ where $\Omega = \Sigma^{-1}$. This result was proved by Sarkar (1969).

Example 2 (Monotone latent variable model)**.** In monotone latent variable model, we assume that the distribution of the P-values $\boldsymbol{P}$ is the marginal distribution of some $(\boldsymbol{P}, \boldsymbol{U})$ where the components of $\boldsymbol{P}$ given $\boldsymbol{U} = \boldsymbol{u}$ are independent and stochastically comonotone in $\boldsymbol{u}$ (Benjamini and Yekutieli, 2001).

Examples of application problems for the above models can be found in Holland and Rosenbaum (1986). Holland and Rosenbaum (1986) also showed that $\boldsymbol{P}$ is positively associated if $\boldsymbol{U}$ is positively associated.

For a given $r \leq n$, the PC null hypothesis and alternative hypothesis are defined as

$$\begin{aligned} H_0^{r/n}: &\quad \{\text{at most } r-1 \text{ hypotheses are non-null}\}, \quad \text{and} \\ H_1^{r/n}: &\quad \{\text{at least } r \text{ hypotheses are non-null}\}. \end{aligned}$$

The null space is defined as $\Theta_0^{r/n} = \{\theta \in \Theta : H_0^{r/n} \text{ is true}\}$.

We use $1{:}r$ to denote $\{1, 2, \ldots, r\}$ and similarly $(r+1){:}n = \{r+1, r+2, \ldots, n\}$. The index set $u \subset 1{:}n$ has cardinality $|u|$ and complement $-u = 1{:}n \setminus u$. Under the null hypothesis H_{0u} we have $\theta_j \in \Theta_{0j}$ for all $j \in u$. The null space of H_{0u} is denoted Θ_{0u}.

Sometimes we combine points $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$ into a point $\boldsymbol{z} \in \mathbb{R}^n$ with $z_j = x_j$ for $j \in u$ and $z_j = y_j$ for $j \notin u$. Such a hybrid point is denoted $\boldsymbol{z} = \boldsymbol{x}_u{:}\boldsymbol{y}_{-u}$. Let $u = \{i_1, i_2, \ldots, i_k\}$, then θ_u is defined as the combination $(\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_k})$.

We can extend the definition of validity to meta-analysis and PC p-values. The combination of k p-values (k may differ from n later) produces the combined p-value $p_{r/k} = f_{r,k}(P_1, \ldots, P_k)$ which is a valid p-value for testing $H_0^{r/k}$ if

$$\sup_{\theta \in \Theta_0^{r/k}} \mathbb{P}_\theta(P_{r/k} \le \alpha) \le \alpha, \quad \forall\, 0 \le \alpha \le 1.$$

Definition 3 (Sensitivity)**.** A sensitive p-value $P_{r/k} = f_{r,k}(P_1, \ldots, P_k)$ for $H_0^{r/k}$ satisfies

$$\limsup_{\boldsymbol{P}_u \to \boldsymbol{0}} P_{r/k} = 0$$

for any $u = \{i_1, i_2, \ldots, i_r\} \subset 1{:}k$ with $|u| = r$, where $\boldsymbol{P}_u = (\boldsymbol{P}_{i_1}, \cdots, \boldsymbol{P}_{i_r})$.

Sensitivity requires that PC p-value drops to 0 when we are certain to reject some subset of r individual hypotheses. We think that it is a practically reasonable requirement for a p-value for testing $H_0^{r/k}$. For a sensitive meta-analysis p-value $P_{1/k}$, we will reject the global null when we are certain to reject at least one of the individual hypotheses.

Here are some examples of valid and sensitive meta-analysis p-values given valid p-values $p_1, \ldots, p_k$. The combination for a method M is defined in terms of a function $f_{M,k}$ which may incorporate sorting of its arguments.

Example 3. Simes' method:

$$p_{S,k} = f_{S,k}(p_1, \ldots, p_k) \equiv \min_{i=1,\cdots,k} \left\{ \frac{k p_{(i)}}{i} \right\}.$$

Simes' P-value is valid when individual p-values satisfy positive regression dependence, which is a special case of being positively associated (Benjamini and Yekutieli, 2001).

Example 4. Fisher's method:

$$p_{F,k} = f_{F,k}(p_1, \ldots, p_k) \equiv \mathbb{P}\biggl(\chi^2_{(2k)} \ge -2 \sum_{i=1}^{k} \log p_i \biggr).$$

Example 5. Weighted Stouffer test: Consider test statistics $T_i \sim \mathcal{N}(\sqrt{n_i}\theta_i/\sigma_i, 1)$, with sample sizes n_i for $i = 1, \ldots, k$ and known $\sigma_i > 0$. The p-value for the null that $\theta_i = 0$ versus the alternative that $\theta_i > 0$ is $p_i = 1 - \Phi(T_i) = \Phi(-T_i)$. A weighted Stouffer p-value for $H_0^{1/k}$ takes the form

$$p_{\mathrm{WS},k} = p_{\mathrm{WS},k}(p_1, \ldots, p_k) \equiv 1 - \Phi\left(\frac{\sum_{i=1}^{k} \sqrt{n_i}\Phi^{-1}(1 - p_i)/\sigma_i}{\sqrt{\sum_{i=1}^{k} n_i/\sigma_i^2}}\right).$$

In fact, $p_{\mathrm{WS},k}$ can also be used for two-sided test as $|\Phi^{-1}(1-p_i)| = O(\sqrt{n_i})$ is also true for two-sided p-values. We shall illustrate the performance under such usage in our simulations in Section 4.

Both Fisher's and the weighted Stouffer's p-values are valid when the individual p-values are independent.

Example 6. Truncated product method (Zaykin et al., 2002): this is a more recently developed method to gain efficiency in the presence of outliers. The test statistic has the form

$$W = \prod_{i:P_i \leq \gamma} P_i$$

where γ is some pre-determined value. The TPM p-value for $H_0^{1/k}$ takes the form

$$p_{\mathrm{TPM},k} = \mathbb{P}(W \leq w_\gamma).$$

where $w_\gamma = \prod_{\{i:P_i \leq \gamma\}} p_i$. The probability function of W was computed for both independent and dependent scenarios in (Zaykin et al., 2002).

For non-symmetric meta-analysis p-values, there are also weighted versions of Fisher tests. Note that each of the functions f in the previous examples is monotone according to this definition:

Definition 4. (Monotonicity) the p-value $f(p_1, \ldots, p_k)$ is monotone if the function f is non-decreasing in each argument. The set of such monotone p-value functions is denoted $\mathcal{F}_{\mathrm{mon}}$. A monotone test is one that rejects its null hypothesis for small values of a monotone p-value.

A non-monotone test would reject its null hypothesis at some input $(p_1, \ldots, p_k)$ but fail to reject at some $(p'_1, \ldots, p'_k)$ with all $p'_i \leq p_i$. Such a test is typically unreasonable. Besides monotonicity, we also clarify the definition of a symmetric combined p-value:

Definition 5. (Symmetry) The combined p-value $f(p_1, \ldots, p_k)$ is symmetric if it equals $f(p_{\pi(1)}, \ldots, p_{\pi(k)})$ where π is any permutation of $\{1, \ldots, k\}$.

Finally, we state the concept of admissibility using the definition of admissible tests from Lehmann and Romano (2006, Chapter 6.7). The hypothesis test of H_0 versus H_1 is described by a function $\varphi(X) \in \{0, 1\}$ of the data X, where $\varphi(X) = 1$ when H_0 is rejected and $\varphi(X) = 0$ otherwise. The test φ is valid at level α if $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\varphi(X)) \leq \alpha$. In our context, the data are a vector $\boldsymbol{P} = (P_1, \dots, P_n)$ of p-values and $\varphi(P_1, \dots, P_n) = 1_{f(P_1,\dots,P_n) \leq \alpha}$ where f is a p-value combination function.

Definition 6 (Ψ,α-admissibility)**.** The level-α test $\varphi \in \Psi$ is α-admissible for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if for any other level-α test $\varphi' \in \Psi$

$$\mathbb{E}_\theta(\varphi') \geq \mathbb{E}_\theta(\varphi), \quad \text{for all } \theta \in \Theta_1$$

implies $\mathbb{E}_\theta(\varphi') = \mathbb{E}_\theta(\varphi)$ for all $\theta \in \Theta_1$.

The definition of admissibility depends on the alternatives in Θ_1 as well as the space Ψ of test functions. The constraints on Θ_1 are important. For ordinary meta-analysis, Birnbaum (1954) shows that every monotone p-value is admissible when the component p-values are independent and the null hypothesis is simple, because there is then some alternative at which that p-value gives optimal power. However, those optimizing alternatives may not all be reasonable. Birnbaum (1955) and Stein (1956) (generalized later by Matthes and Truax (1967) to include nuisance parameters) also showed that for the ordinary meta-analysis, when the test statistic distribution is an exponential family with θ as canonical parameter, a necessary and sufficient condition for admissibility is to have a closed convex acceptance region of underlying test statistics.

The space of test functions Ψ, traditionally contains all possible functions when considering admissibility. However, for PC tests we restrict Ψ to only include tests using monotone p-values . This lets us avoid some unreasonable but more powerful tests (see Section 3.2 for details).

2.2 BHPC p-values

Now we restate a version of Theorem 1 from Benjamini and Heller (2008).

Theorem 2.1. *Let $P_1, \dots, P_n$ be independent valid p-values, and for $k = n-r+1$ let $f_{M,k}(P_1, \dots, P_k)$ be a valid and symmetric meta-analysis p-value where $f_{M,k} \in \mathcal{F}_{\text{mon}}$. Then $P_{r/n} = f_{M,n-r+1}(P_{(r)}, P_{(r+1)}, \dots, P_{(n)})$ is a valid p-value for $H_0^{r/n}$.*

Remark 2.1. The original result in Benjamini and Heller (2008) is more general. It includes some dependent p-value cases when $f_{M,k}$ is the Sime's or Bonferroni p-value.

As mentioned, we call the combined p-value $P_{r/n}$ described in Theorem 2.1 a BHPC p-value for short. In practice it makes sense to require that the p-value combination function $f_{M,k}(\cdot)$, for $k = n - r + 1$, be a sensitive one for $H_0^{1/k}$. Notice that if $f_{M,k}$ were a partial conjunction test of $H_0^{s/k}$ for $s > 1$, then $f_{M,k}$ is still a valid meta-analysis p-value but is not sensitive any more. The BHPC p-values satisfy a nesting property in that $P_{r/n}$ in Theorem 2.1 also yields a valid test of $H_0^{(r+s-1)/n}$. While valid, that test would have less power than one based on $P_{(r+s-1)/n}$.

Based on the relationship between hypotheses testing and building confidence set, if we have for each $r = 1, 2, \dots, n$ a valid combined p-value $p_{r/n}$ for $H_0^{r/n}$, then the $1 - \alpha$ confidence set for the true number of non-null hypotheses is

$$I = \{r : P_{r/n} \le \alpha\}.$$

Benjamini and Heller (2008) showed that if each $P_{r/n}$ is a BHPC p-value with $f_{M,k} = f_{S,k}$ in Example 3, then $P_{1/n} \le P_{2/n} \le \cdots \le P_{n/n}$ and $I = [\hat{r}, n]$ with $\hat{r} = \max\{r : P_{r/n} \le \alpha\}$ becomes a $1 - \alpha$ confidence interval for the true number of non-null hypotheses.

3 GBHPC p-values

Motivated by the BHPC p-value, we discuss a more general class of combined p-values with good power properties. These are the GBHPC p-values:

Definition 7 (GBHPC p-value)**.** For each $u \subset 1{:}n$ with $|u| = n - r + 1$, let g_u be a function from $[0,1]^{|u|}$ to $[0,1]$ that is non-decreasing in each component and for which g_u provides a valid meta-analysis p-value for H_{0u}. Then

$$f^{\star}(\boldsymbol{p}) = f^{\star}(p_1, \cdots, p_n) \equiv \max_{\substack{u \subset 1:n \\ |u| = n-r+1}} g_u(\boldsymbol{p}_u) \tag{3}$$

is a generalized BHPC (GBHPC) p-value for $H_0^{r/n}$.

The alternative hypothesis $H_1^{r/n}$ that at least r out of n hypotheses are false is equivalent to the statement that for every $n - r + 1$ of the n hypotheses, there is at least one of them that is false. This is the reason that the GBHPC p-value is the maximum of all the meta-analysis p-values of size $n - r + 1$. Using above explanation, the next proposition states validity of GBHPC p-values under any dependence structure of the individual p-values, which is a simple result also mentioned in Benjamini et al. (2009).

Proposition 3.1. *Any GBHPC p-value is a valid p-value for $H_0^{r/n}$.*

Proof. Consider a GBHPC p-value of the form (3). From the definition of $H_0^{r/n}$, for all $\theta \in \Theta_0^{r/n}$, there exists u with $|u| = n - r + 1$ such that $\theta_j \in \Theta_{0j}$ for all $j \in u$. then for any $\alpha \in [0, 1]$,

$$\mathbb{P}_\theta(f^\star(\boldsymbol{P}) \leq \alpha) \leq \mathbb{P}_\theta(g_u(\boldsymbol{P}_u) \leq \alpha) = \mathbb{P}_{\theta_u}(g_u(\boldsymbol{P}_u) \leq \alpha) \leq \alpha.$$

Thus $f^\star(P_1, \cdots, P_n)$ is valid for $H_0^{r/n}$. □

The BHPC p-value is a special case of GBHPC p-values. If all $g_u \equiv g$ in (3) and g a monotone and symmetric combined p-value, then $f^\star(\boldsymbol{p}) = g(p_{(r)}, \ldots, p_{(n)})$ becomes a BHPC p-value. Some meta-analysis methods, such as the weighted Stouffer test in Example 5, treat their component p-values differently depending on the relative sample sizes on which they are based. The GBHPC framework includes such methods.

Next we show that a sensitive GBHPC p-value has a unique representation in the form of (3).

Proposition 3.2. *If a GBHPC p-value $f^\star(P_1, \ldots, P_n)$ for $H_0^{r/n}$ is sensitive, then each g_u is also sensitive and the representation of $f^\star$ in* (3) *is unique with*

$$g_u(\boldsymbol{P}_u) = \inf_{\boldsymbol{P}_{-u} \in (0,1]^{r-1}} f^\star(P_1, \cdots, P_n). \tag{4}$$

Proof. Consider any given $u \subset 1{:}n$ with $|u| = n - r + 1$. Then for any $i \in u$, let $u_i = -u \cup \{i\}$, so $|u_i| = r$. As $f^\star$ is sensitive, we have

$$\liminf_{P_i \to 0} g_u(\boldsymbol{P}_u) = \liminf_{\boldsymbol{P}_{u_i} \to \boldsymbol{0}} g_u(\boldsymbol{P}_u) \leq \liminf_{\boldsymbol{P}_{u_i} \to \boldsymbol{0}} f^\star(\boldsymbol{P}) = 0$$

Thus, g_u is sensitive for every u with $|u| = n - r + 1$. On the other hand, for a given u_0 with $|u_0| = n - r + 1$, using definition (3), we have

$$\inf_{\boldsymbol{P}_{-u_0} \in (0,1]^{r-1}} f^\star(\boldsymbol{P}) = \inf_{\boldsymbol{P}_{-u_0} \in (0,1]^{r-1}} \left[\max_u g_u(\boldsymbol{P}_u)\right] = g_{u_0}(\boldsymbol{P}_{u_0})$$

which proves the uniqueness of the representation. □

Remark 3.1. Equation (4) also shows that any symmetric GBHPC p-value is a BHPC p-value. If $f^\star$ is symmetric, then g_u constructed in (4) is also valid, monotone and symmetric. Also, $g_u \equiv g$ is the same for all u. Thus $f^\star = g(P_{(n)}, \cdots, P_{(r)})$ is a BHPC p-value.

Remark 3.2. Equation (3) involves taking the maximum of meta-analysis p-values over $\binom{n}{r-1}$ distinct subsets. This could be costly but if $r = 2$ or $r = n - 1$ then $\binom{n}{r-1} = n$, which is manageable. It is sometimes possible to use the special structure of the problem to construct easy-to-compute non-symmetric GBHPC p-values. For instance in Section 5, our GBHPC p-values for the pharmaceutical data have almost the same computational cost as BHPC p-values, but can give much smaller p-values.

3.1 Monotone α-admissiblity

Now we discuss the sufficient and necessary conditions for admissible combined p-values of a PC hypothesis. Each of our results uses some combination of the following three assumptions on the individual p-values.

Assumption 1 (Strong alternatives)**.** $\forall \alpha > 0$ and $i = 1, \dots, n$, $\sup_{\theta_i \in \Theta_{1i}} \mathbb{P}_{\theta_i}(P_i \le \alpha) = 1$.

Assumption 2 (Continuity)**.** For any $\theta_i \in \Theta_{1i}$ and $i = 1, \dots, n$, $\mathbb{P}_{\theta_i}(P_i = 0) = 0$.

Assumption 3 (Completeness)**.** The family $\{\mathbb{P}_{\theta_u} : \theta_u \notin \Theta_{0u}\}$ for any subset u with $|u| = n - r + 1$ is complete.

Assumption 1 states that for each individual hypothesis there are strong enough alternatives for which we can almost certainly reject the null. Assumption 2 is a technical condition under which we never get a p-value that is exactly zero. That is, there is never absolute certainty. The completeness in Assumption 3 is to guarantee that if two level α meta-analysis tests for H_{0u} have the same power at every point in the alternative space, then they are the same test. Roughly speaking, both Assumptions 1 and 3 require that the alternative space of each individual hypothesis is large enough to include various possibilities.

All three assumptions are satisfied in common tests. Tests satisfying Assumption 1 include tests for the parameters of exponential families and location families. Lehmann and Romano (2006, Theorem 4.3.1) show that completeness is satisfied when testing the natural parameter of a k-dimensional exponential family, if the alternative space Θ_{1i} contains a k-dimensional rectangle. If the individual p-values are independent, then completeness of the alternative space for each individual hypothesis implies Assumption 3. Thus, we believe that assumptions 1 to 3 cover a large class of problems.

Theorem 3.3 shows that GBHPC p-values form a complete class of monotone α-admissible p-values for $H_0^{r/n}$. Theorem 3.4 states that a sufficient condition for a sensitive GBHPC p-value to be monotone α-admissible is that each g_u is admissible for H_{0u}.

Theorem 3.3. *Let $P_1, \dots, P_n$ be positively associated P-values satisfying assumptions 1 and 2. Let $P_{r/n}$ be a valid monotone p-value for $H_0^{r/n}$. Then there exists a valid GBHPC p-value $p^\star_{r/n}$ that is uniformly at least as powerful as $P_{r/n}$.*

Theorem 3.4. *Let $P_1, \dots, P_n$ be positively associated P-values satisfying assumptions 1 to 3. For a sensitive GBHPC p-value $P^\star_{r/n} = f^\star(\boldsymbol{P})$ of the form (3), a sufficient condition for $P^\star_{r/n}$ to be monotone α-admissible is that each g_u is an admissible meta-analysis p-value for H_{0u}.*

Before proving Theorems 3.3 and 3.4 we introduce the following lemma, which underpins them both. It shows that given a valid monotone p-value that is not of the GBHPC form, we can expand its rejection region while retaining its validity.

Lemma 3.5. *Let $P_1, \dots, P_n$ be positively associated P-values satisfying Assumption 1. Let $f(P_1, \cdots, P_n)$ be a valid monotone p-value for $H_0^{r/n}$ and for $u \subset 1{:}n$ with $|u| = n - r + 1$, define*

$$g_u(\boldsymbol{P}_u) = \inf_{\boldsymbol{P}_{-u} \in (0,1]^{r-1}} f(P_1, \cdots, P_n). \tag{5}$$

Then g_u is a valid monotone meta-analysis p-value for H_{0u}.

Proof. Monotonicity of f implies monotonicity and measurability of g_u. Next, suppose that g_u is not valid for H_{0u}. Then there is an $\alpha \in [0,1]$ and a $\theta^\star_u$ with $\theta_j \in \Theta_{0j}$ for all $j \in u$ such that $\mathbb{P}_{\theta^\star_u}(g_u(\boldsymbol{P}_u) \le \alpha) = \mathbb{P}_{\theta^\star_u}\big(\inf_{\boldsymbol{P}_{-u} \in (0,1]^{r-1}} f(\boldsymbol{P}) \le \alpha\big) > \alpha + \epsilon$ for some $\epsilon > 0$. From the monotonity of f, there is some fixed $\widetilde{p} \in (0,1]$ with $\mathbb{P}_{\theta^\star_u}(f(\boldsymbol{P}_u{:}\boldsymbol{p}_{-u}) \le \alpha) > \alpha + \epsilon$ for any $\boldsymbol{p}_{-u} \in [0, \widetilde{p}]^{r-1}$. Since the p-values are positively associated and $\{\boldsymbol{p} : f(\boldsymbol{p}_u, \boldsymbol{p}_{-u}) \le \alpha\}$ is a decreasing set for any fixed $\boldsymbol{p}_{-u}$, using (2) we have

$$\mathbb{P}_\theta\left(f(\boldsymbol{P}_u{:}\boldsymbol{p}_{-u}) \le \alpha \mid \boldsymbol{P}_{-u} \in [0, \widetilde{p}]^{r-1}\right) \ge \mathbb{P}_\theta\left(f(\boldsymbol{P}_u{:}p_{-u}) \le \alpha\right) \ge \alpha + \epsilon$$

for any θ satisfying $\theta_u = \theta^\star_u$ and any fixed $\boldsymbol{p}_{-u} \in [0, \widetilde{p}]^{r-1}$. Therefore

$$\mathbb{P}_\theta\left(f(\boldsymbol{P}) \le \alpha \mid \boldsymbol{P}_{-u} \in [0, \widetilde{p}]^{r-1}\right) \ge \alpha + \epsilon.$$

Using Assumption 1, there also exists $\theta^\star_{-u}$ with $\theta^\star_j \in \Theta_{1j}$ for any $j \in -u$ such that $\mathbb{P}_{\theta^\star_j}(P_j \le \widetilde{p}) \ge \big((\alpha + \epsilon/2)/(\alpha + \epsilon)\big)^{1/(r-1)}$. Since the p-values are positively associated, using (1) we have

$$\mathbb{P}_{(\theta^\star_u : \theta^\star_{-u})}\big(f(\boldsymbol{P}) \le \alpha\big) \ge \mathbb{P}_{(\theta^\star_u : \theta^\star_{-u})}\Big(f(\boldsymbol{P}) \le \alpha, P_j \le \widetilde{p}, \forall j \in -u\Big) > \alpha + \epsilon/2$$

contradicting the validity of $f(\boldsymbol{P})$. □

Now we are ready to prove Theorems 3.3 and 3.4.

Proof of Theorem 3.3. Let $g_u(\boldsymbol{P}_u)$ be defined in (5). Then $P_{r/n} \geq P^\star_{r/n}$ when $\boldsymbol{P} \in (0,1]^n$. Using Assumption 2, $P^\star_{r/n}$ is then uniformly at least as powerful as $P_{r/n}$. It then follows directly from Lemmas 3.1 and 3.5 that $p^\star_{r/n}$ is a valid GBHPC p-value. □

Using Lemma 3.1, to prove Theorem 3.4, we only need to prove the monotone α-admissibility of $P^\star_{r/n}$.

Proof of Theorem 3.4. To prove the monotone α-admissibility of $f^\star(P_1, \cdots, P_n)$, suppose that there is a valid monotone test $f^{\star\star}$ satisfying $\mathbb{P}_\theta(f^{\star\star}(P) \leq \alpha) \geq \mathbb{P}_\theta(f^\star(P) \leq \alpha)$ for all $\theta \in \Theta_1^{r/n}$. By Theorem 3.3 we can assume that $f^{\star\star}$ is a GBHPC p-value:

$$f^{\star\star}(\boldsymbol{P}) = \max_{\substack{u \subset 1:n \\ |u| = n-r+1}} g'_u(\boldsymbol{P}_u),$$

where g'_u is a valid monotone meta-analysis p-value. Notice that since $f^\star$ is sensitive, equation (4) holds. We now show that for each $u \subset 1{:}n$ with $|u| = n - r + 1$, and any $\theta_u \notin \Theta_{0u}$,

$$\mathbb{P}_{\theta_u}\Big(\inf_{\boldsymbol{P}_{-u} \in (0,1]^{r-1}} f^\star(\boldsymbol{P}) \leq \alpha \Big) \leq \mathbb{P}_{\theta_u}(g'_u(\boldsymbol{P}_u) \leq \alpha) \equiv \beta' \tag{6}$$

using a similar strategy as in the proof of Lemma 3.5. If (6) does not hold for some set u and a corresponding θ_u, then there exists some $\epsilon > 0$ and $\widetilde{p} \in (0,1]$ such that $\mathbb{P}_{\theta_u}(f^\star(\boldsymbol{P}_u{:}\boldsymbol{p}_{-u}) \leq \alpha) > \beta' + \epsilon$ for any $\boldsymbol{p}_{-u} \in (0, \widetilde{p}]^{r-1}$. Using Assumption 1, there exists $\theta^\star$ with $\theta^\star_j \in \Theta_{1j}$ for $j \in -u$ such that $\mathbb{P}_{\theta^\star_j}(P_j \leq \widetilde{p}) \geq \big((\beta' + \epsilon/2)/(\beta' + \epsilon)\big)^{1/(r-1)}$. Thus,

$$\begin{aligned} \mathbb{P}_{(\theta_u:\theta^\star_{-u})}(f^\star(\boldsymbol{P}) \leq \alpha) &\geq \mathbb{P}_{(\theta_u:\theta^\star_{-u})}(f^\star(\boldsymbol{P}) \leq \alpha, P_j \leq \widetilde{p}, \forall j \in -u) > \beta' + \epsilon/2 \\ &> \mathbb{P}_{\theta_u}(g'_u(\boldsymbol{P}_u) \leq \alpha) \geq \mathbb{P}_{(\theta_u:\theta^\star_{-u})}(f^{\star\star}(\boldsymbol{P}) \leq \alpha) \end{aligned}$$

which violates the assumption that $f^{\star\star}$ is uniformly at least as powerful as $f^\star$. Thus, equation (6) holds. Now equations (4) and (6) imply that $\mathbb{P}_{\theta_u}(g'_u \leq \alpha) \geq \mathbb{P}_{\theta_u}(g_u(\boldsymbol{P}_u) \leq \alpha)$ for any $\theta_u \notin \Theta_{0u}$ and any $\alpha \in [0,1]$. As $g_u(\boldsymbol{P}_u)$ is α-admissible for H_{0u}, we have $\mathbb{P}_{\theta_u}(g'_u(\boldsymbol{P}_u) \leq \alpha) = \mathbb{P}_{\theta_u}(g_u(\boldsymbol{P}_u) \leq \alpha)$. Further, using Assumption 3 we have $g'_u(\boldsymbol{P}_u) = g_u(\boldsymbol{P}_u)$ a.e.. Thus, for all $\theta \in \Theta_1^{r/n}$, $\mathbb{P}_\theta(f^{\star\star}(\boldsymbol{P}) \leq \alpha) = \mathbb{P}_\theta(f^\star(\boldsymbol{P}) \leq \alpha)$ which shows that $f^\star$ is monotone α-admissible for $H_0^{r/n}$. □

Remark 3.3. As mentioned in Remark 3.1, the BHPC p-values are symmetric GBHPC p-values As a consequence, the BHPC p-values characterize the form of symmetric monotone admissible combined p-values.

Combining Theorem 3.4 with results of Birnbaum (1955) and Lehmann and Romano (2006, Theorem 6.7.1) who characterized admissible tests for the global null in exponential families, we can give more specific conditions when Theorem 3.4 is applied to exponential families. Here is the result for a simple scenario.

Example 7. Suppose that independent test statistics T_i for $i = 1, \ldots, n$ are available on hypotheses $H_{0i} : \boldsymbol{\theta}_i = \boldsymbol{a}_i \in \mathbb{R}^{k_i}$ against $H_{0i} : \boldsymbol{\theta}_i \in \mathbb{R}^{k_i} \setminus \{\boldsymbol{a}_i\}$. Here we assume that every T_i is the sufficient statistic for an exponential family with natural parameter $\boldsymbol{\theta}_i$. For a sensitive GBHPC p-value $f^\star(\boldsymbol{p})$, suppose that $\mathbb{P}_{\boldsymbol{\theta}_{0u}}(g_u(\boldsymbol{P}_u) \leq \alpha) = \alpha$. Also, for any $\alpha \in [0, 1]$ the set of $\boldsymbol{T}_u$ for which $g_u(\boldsymbol{p}_u) > \alpha$ (the acceptance region) is a closed and convex set, except for a subset of measure 0. Then $f^\star(\boldsymbol{p})$ is monotone α-admissible for $H_0^{r/n}$.

Related work on convexity and admissibility also appears in Matthes and Truax (1967) for testing parameters of exponential families with presence of nuisance parameters, Marden (1982) and Brown and Marden (1989) for generalization to distribution families beyond exponential families, and Owen (2009) for tests powerful against alternatives with concordant signs. Notice that the n-dimensional set of test statistics $\boldsymbol{T}$ itself for which $f^\star(\boldsymbol{p}) > \alpha$ is not convex. For partial conjunctions, the null hypothesis for the parameter usually includes all of the coordinate axes and the smallest convex set containing the axes is all of Euclidean space. As a result, convexity of the acceptance region cannot be an appropriate admissibility criterion for partial conjunction testing.

3.2 Inadmissibility

In Section 3.1 we constructed monotone α-admissible p-values for $H_0^{r/n}$, we show that they fail to be admissible if we allow non-monotone tests. For the case $n = r = 2$, the construction of such counter-examples dates back to Lehmann (1952) and Iwasa (1991).

Here we demonstrate that if we do not require monotone tests, then a BPHC test is inadmissible. Let $n = r = 2$. If both P_1 and P_2 are α-admissible, then using Theorems 3.3 and 3.4, the constructed combined p-value is just $P_{(2)}$, which is monotone admissible. At a given α, the critical function is $\varphi = \mathbb{1}_{p_{(2)} \leq \alpha}(p_1, p_2)$.

Now we can easily construct a more powerful α-level test, by adding to the original rejection region a square around the top-right corner in the p-value space (solid shaded regions in Figure 1). Define the set

$$S = \begin{cases} \{(p_1, p_2) \mid p_{(1)} \geq 1 - \alpha\}, & \text{if } \alpha < \frac{1}{2} \\ \{(p_1, p_2) \mid p_{(1)} \geq \alpha\}, & \text{if } \alpha \geq \frac{1}{2}. \end{cases}$$

Then the test φ' with critical function $\varphi'(\boldsymbol{P}) = \varphi(\boldsymbol{P}) + \mathbb{1}_{(P_1,P_2)\in S}$ is uniformly and strictly more powerful than φ. To prove that φ' is an α-level test, we note that $S \cap \{p_{(2)} \leq \alpha\} = \varnothing$. Therefore $\mathbb{E}(\varphi'(\boldsymbol{P}) \mid P_1 = p_0) \leq \alpha$ holds for any $p_0 \in [0, 1]$. Similarly, $\mathbb{E}(\varphi'(\boldsymbol{P}) \mid P_2 = p_0) \leq \alpha$. Since p_0 is arbitrary we conclude that φ' is an α-level test. Actually, as shown in Figure 1, we can further expand the rejection region of φ' to include also the dotted shaded regions and to get an even more powerful but still valid test $\widetilde{\varphi}$. The rejection region of $\widetilde{\varphi}$ in the p-values space consists of small squares along the diagonal line.

If the test statistics are $Z_1 \sim \mathcal{N}(\mu_1, 1)$ and $Z_2 \sim \mathcal{N}(\mu_2, 1)$, and H_1 and H_2 are two-sided tests for the mean μ_1 and μ_2 respectively, then the top two plots of Figure 1 show the rejection region of φ' and $\widetilde{\varphi}$ at level $\alpha = 0.1$ in the p-value space and in the test statistic space. The bottom two plots compare the power of φ and $\widetilde{\varphi}$ as a function of (μ_1, μ_2). They show that the power gain of the non-monotone $\widetilde{\varphi}$ only appears in the low power region where the power is below or near α.

The more powerful test φ' increases power by strangely rejecting $H_0^{2/2}$ when $\max(p_1, p_2)$ is large enough. We now use this same approach to show that without the monotonicity constraint, any GBHPC p-value is inadmissible for any n and any $r \in 2{:}n$. The counter-examples reject $H_0^{r/n}$ when all p-values are large. The idea is to show that for any GBHPC test, it's always possible to add a "box"-shaped rejection region like the square around the origin in the right panel of Figure 1 while still keeping the test valid. The point is not to advocate for such tests, but rather to reinforce the idea that admissibility is only a useful concept within a well chosen class of functions.

We need the following mild technical constraint to guarantee that the "box" we choose can really increase power at least in one alternative hypothesis.

Assumption 4. For each $i \in 1{:}n$, there exists $\theta_i^0 \in \Theta_{0i}$ that $\mathbb{P}_{\theta_i^0}(P_i \leq \alpha) = \sup_{\theta_i \in \Theta_{0i}}(P_i \leq \alpha)$ for any $\alpha \in [0, 1]$. Let $\theta^0 = (\theta_1^0, \theta_2^0, \cdots, \theta_n^0)$. Then for any set A, if $\mathbb{P}_{\theta^0}(A) > 0$, then there exists $\theta^1 \in \Theta_1^{r/n}$ that $\mathbb{P}_{\theta^1}(A) > 0$.

Theorem 3.6. *Let $P_1, \ldots, P_n$ be independent p-values satisfying Assumptions 1, 2 and 4. Let $1 < r \leq n$ and $\alpha \in (0, 1)$. Then any monotone α-admissible combined p-value for testing $H_0^{r/n}$ is not α-admissible without the monotonicity constraint.*

Proof. Using Theorem 3.3, we only need to consider a GBHPC p-value $f^\star$ which is defined in Definition 7. Let $\theta^0 = (\theta_1^0, \theta_2^0, \cdots, \theta_n^0)$ be the parameter in Assumption 4. define

$$R = \{\boldsymbol{p} \in [0, 1]^n : f^\star(\boldsymbol{p}) \leq \alpha\} = \bigcap_{\substack{u \subset 1:n \\ |u| = n-r+1}} R_u$$

where $R_u = \{\boldsymbol{p} \in [0, 1]^n : g_u(\boldsymbol{p}_u) \leq \alpha\}$ and g_u is defined in (3).

First, as $\mathbb{P}_{\theta^0}(f^\star \leq \alpha) \leq \alpha < 1$ and $f^\star$ is non-decreasing, there exists some $p_0 < 1$ such that if $p_j \geq p_0$ for all $j \in 1{:}n$ then $f^\star(\boldsymbol{p}) > \alpha$.

Then, we show that there must exist a set $u^\star$ with $\mathbb{P}_{\theta^0}(R_{u^\star} \cap R^c) = \epsilon > 0$, where R^c is the complement set of R. If this doesn't hold, then it means that for any $u \subset 1{:}n$ with $|u| = n - r + 1$, the equation $1_{f^\star(\boldsymbol{p}) \leq \alpha}(\boldsymbol{p}) = 1_{g_u(\boldsymbol{p}_u) \leq \alpha}(\boldsymbol{p})$ *a.e.* $\mathbb{P}_{\theta^0}$ holds. This implies that $1_{f^\star \leq \alpha}$ doesn't depend on $\boldsymbol{p}_{-u}$ except for a zero probability set under $\mathbb{P}_{\theta^0}$. As $\cup_{\substack{u \subset 1:n \\ |u| = n-r+1}} -u = 1{:}n$, we get that $1_{f^\star \leq \alpha}$ doesn't depend on any p_j except for a zero probability set under $\mathbb{P}_{\theta^0}$, which implies that $1_{f^\star \leq \alpha} \equiv 1$ or 0 *a.e.* $\mathbb{P}_{\theta^0}$. It is obvious that such a test is either invalid or trivially not admissible, which contradicts our assumptions.

As a consequence, we have $\mathbb{P}_{\theta^0}(f^\star \leq \alpha) = \mathbb{P}_{\theta_0}(R_{u^\star}) - \epsilon \leq \alpha - \epsilon$. Notice that $\mathbb{P}_{\theta^0}(f^\star \leq \alpha) = \mathbb{E}_{\theta^0_{-u}}\big(\mathbb{P}_{\theta^0_u}[f^\star \leq \alpha \mid \boldsymbol{P}_{-u}]\big)$ for any u. Using the fact that $f^\star$ is non-decreasing, $\mathbb{P}_{\theta^0_u}[f^\star \leq \alpha \mid \boldsymbol{P}_{-u} = \boldsymbol{p}_{-u}]$ is non-increasing in $\boldsymbol{p}_{-u}$. Thus there exists $\widetilde{p} < 1$, such that for any u, if $\boldsymbol{p}_{-u} \in [\widetilde{p}, 1]^{r-1}$, then

$$\mathbb{P}_{\theta^0_u}[f^\star \leq \alpha \mid \boldsymbol{P}_{-u} = \boldsymbol{p}_{-u}] \leq \alpha - \epsilon.$$

Let $p^\star = \max(p_0, \widetilde{p}, 1 - \epsilon^{1/(n-r+1)})$ and $S = \cap_i\{\boldsymbol{p} \in [0,1]^n : p_i \geq p^\star\}$. Then we construct a new test with critical function φ: $\varphi = \mathbb{1}_{f^\star \leq \alpha} + \mathbb{1}_S$.

As $\{\boldsymbol{p} \in [0,1]^n : f^\star(\boldsymbol{p}) \leq \alpha\} \cap S = \varnothing$, we know that φ is at least as powerful as $\mathbb{1}_{f^\star \leq \alpha}$. Using Assumption 4, as $\mathbb{P}_{\theta^0}(S) \geq (1 - p^\star)^n > 0$, there exists $\theta^1 \in \Theta_1^{r/n}$ with $\mathbb{P}_{\theta^1}(S) > 0$. Thus, φ strictly dominates $\mathbb{1}_{f^\star \leq \alpha}$ at θ^1. Finally, for any $\boldsymbol{p} \in [0,1]^n$ and any $u \subset 1{:}n$ with $|u| = n - r + 1$, if $\theta_u \in H^{1/n-r+1}$, then

$$\begin{aligned}
\mathbb{E}_{\theta_u}[\varphi \mid \boldsymbol{P}_{-u} = \boldsymbol{p}_{-u}] &\leq \mathbb{P}_{\theta_u}[f^\star \leq \alpha \mid \boldsymbol{P}_{-u} = \boldsymbol{p}_{-u}] + \epsilon \mathbb{1}_{\boldsymbol{p}_{-u} \in [p^\star, 1]^{r-1}} \\
&\leq \mathbb{P}_{\theta^0_u}[f^\star \leq \alpha \mid \boldsymbol{P}_{-u} = \boldsymbol{p}_{-u}] + \epsilon \mathbb{1}_{\boldsymbol{p}_{-u} \in [p^\star, 1]^{r-1}} \leq \alpha.
\end{aligned}$$

The second inequality above follows from Assumption 4, independence of the individual p-values and monotonicity of $f^\star$. Thus φ is still an α-level test for $H^{r/n}$. This shows that $f^\star$ is not α-admissible. □

4 Simulation

In this simulation example, we compare the power of several GBHPC p-values testing for the PC hypothesis $H_0^{r/n}$ with $n = 8$ studies and $r = 2$. Compared with other r values, the null hypothesis $H_0^{2/n}$ is often of particular interest as it tests whether the significance of the effect replicates at least once across studies. It is also a case where the number of tests computed in a non-symmetric GBHPC p-value is at most n.

We consider alternatives whose true number of non-null hypotheses r_0 is 2, 4, or 6. We assume that all the individual test statistics are independent. Each p-value P_i is a two-sided p-value of the corresponding z-value $Z_i \sim \mathcal{N}(\sqrt{N_i}\mu_i, 1)$ for $i = 1, 2, \ldots, 8$. We set three of the sample sizes N_i of the eight individual studies to 100, another three of them to 500 and the last two to 1000. If H_{0i} is true, then $\mu_i = 0$. When H_{0i} is false, we generate $\mu_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_0, \beta_0)$, for a range of (α_0, β_0) pairs. We define α_0 and β_0 in terms of $\mu_0 = \alpha_0/\beta_0$ and $\sigma_0 = \sqrt{\alpha_0/\beta_0^2}$ which are the mean and standard deviation of the non-null effect across studies. We compare the power of each GBHPC p-value as a function of (μ_0, σ_0) at the significance level of $\alpha = 0.05$.

We compare three GBHPC P-values, whose meta-analysis P-values g_u are from examples 3 to 5 respectively. The results are shown in Figure 2. For each r_0 and each form of the GBHPC p-value, we plot a map of the power of Fisher's test versus μ_0 and σ_0. For Simes and weighted Stouffer GBHPC p-values, we plot their powers after subtracting the power of Fisher's BHPC p-value at the corresponding location. Here are some observations from the simulation results. First, Figure 2 shows that when $r_0 = 2$, Simes BHPC p-value is the most powerful in a large region of the alternative space. The reason is that, for each subset of hypotheses with size $n - r_0 + 1 = n - 1$, at the worst case there is only one non-null individual hypothesis, where Simes should be most powerful in detecting extreme p-values. Second, the Weighted Stouffer GBHPC p-value can have higher power than Fisher's BHPC p-value when $r_0 > r = 2$ and when the effect heterogeneity across non-null studies (σ_0) is not too large to dominate the average effect (μ_0). Notice that we are using two-sided p-values to make a fair comparison of the methods, thus taking $\sqrt{n_i}$ as weights in Stouffer's method would not be optimal. However, as we have discussed in Example 5 and shown here, it can still provide a good test with two-sided p-value. If individual p-values were one-sided, then we would have seen an even larger power gain. Finally, the three methods do not have a noticeable difference in power when the power is very low or very high. Most of the difference appears when the power is in the range from 0.4 to 0.6.

5 Anticoagulant data

Compared with BHPC p-values, GBHPC p-values has the flexiblility to make use of the possibly complex dependence structure across studies, and thus may achieve higher power. We use a real data example to illustrate this benefit.

The dataset (Ruff et al., 2014) is a pooled dataset from four randomized clinical trials aiming to measure the relative benefit of new oral anticoagulants (NOAC) compared with an older anticoagulant (warfarin) for stroke prevention. One primary goal in the original paper was to assess and compare the efficiency of these new drugs in different clinical subgroups of patients. Eighteen subgroups are shown in Table 2a. The data for each one is a pair (m, N) where N is the number of people in the subgroup, and of those people, m suffer a stroke or systemic embolic event. We have (m, N) for both warfarin and NOAC patients.

The N people in each subgroup, such as people with low creatinine clearance, are combined from all four of the underlying studies. Instead of doing a PC test about how many of the four studies have significant differences, we will do PC tests on the number of these clinically relevant subgroups that have different outcomes. Each subgroup p-value is based on Fisher's exact test for an equal odds ratio between the warfarin and NOAC subjects in that subgroup. We are ignoring any heterogeneity among subjects from the four different clinical trials because our purpose is simply to illustrate GBHPC tests.

One major difficulty in applying PC tests (and meta-analysis) to these subgroups is that they have substantial overlap. For example, the subgroup of age < 75 clearly overlaps with the subgroup of female. As a result, the test statistics are dependent in a complex way that is not known to us from the given data. When there is unknown dependence across individual hypotheses, the only BHPC p-values available take the meta-analysis p-value g to be either the Bonferroni or Simes p-value. The left column of Table 2b has the values of the Bonferroni BHPC p-values $p_{r/n}^{\mathrm{Bon}}$ for $H_0^{r/n}$ with $n = 18$ and r ranging from 2 to 18.

The Bonferroni or even the Simes PC values might be too conservative. Some of the 18 subgroups have disjoint sets of subjects, so we could model their test statistics as independent of each other. For example, the three p-values for the three CHADS_2 groups are independent. Our strategy is to use Fisher's method within each grouping factor and Bonferroni between factors. This will provide a valid GBHPC p-value.

In more detail, let $1{:}18 = \cup_{i=1}^{8} I_i$ where each I_i is the index set of subgroups defined by the i'th grouping factor. To build a GBHPC p-value $p_{r/n}^{\mathrm{new}}$, we construct $g_u(p_u)$ for each $u \subset 1{:}n$ with $|u| = n - r + 1$ as follows. Let $v = v(u) = \{i \mid I_i \cap u \neq \varnothing\}$. For each $i \in v$, a Fisher's p-value (Example 4) $p_{u,i}$ is calculated on $p_{u \cap I_i}$, then

$$g_u(p_u) = |v| \cdot \min_{i \in v(u)} p_{u,i},$$

which is a Bonferroni combination of p-values across the grouping factors.

Next we consider computation of $p_{r/n}^{\mathrm{new}}$ for $r = 2, 3, \ldots, n$. That amounts to $2^n - n - 1$ tests. We can however speed up computation by checking Table 2c. Table 2c has all $p_{u,i}$ values that can possibly affect the value of $p_{r/n}^{\mathrm{new}}$ for any r. Notice that the GBHPC p-value is

$$p_{r/n}^{\mathrm{new}} = \max_{u, |u| = n-r+1} \left\{ |v| \cdot \min_{i \in v(u)} p_{u,i} \right\}.$$

Since $p_{u,i}$ is symmetric, it can possibly influence $f^\star$ only when $p_{u,i}$ is the Fisher's combination on the largest $|u \cap I_i|$ p-values in I_i. Specifically, we define

$$A_{s,i} = \max_{|u \cap I_i| = s, u \subset 1:n} p_{u,i} \tag{7}$$

and set $A_{0,i} = 1$ for $i = 1, 2, \cdots, 8$. Table 2c lists all $A_{s,i}$ values for our data. Now the GBHPC p-value $p_{r/n}^{\text{new}}$ can be rewritten as

$$p_{r/n}^{\text{new}} = \max_{\sum_{i=1}^{8} s_i = n-r+1} \left\{ \left(\sum_i 1_{s_i > 0} \right) \cdot \min_i A_{s_i,i} \right\} \tag{8}$$

To compute all $p_{r/n}^{\text{new}}$ using (8), we need at most n^2 tests, which can be much smaller than $2^n - n - 1$. To see this, we first rank all $A_{s_i,i}$ from the smallest to the largest and start from the smallest $A_{s_i,i}$. Given that $\min_i A_{s_i,i}$ is taken at some fixed $\{s_{i_0}, i_0\}$, we can quickly find the largest possible s_i for each $i \neq i_0$ as we start from the smallest $A_{s_i,i}$. Thus, for each r, we can compute the largest $\sum_i 1_{s_i>0}$ given that $\min_i A_{s_i,i}$ is taken at the fixed $\{s_{i_0}, i_0\}$ and $\sum_i s_i = n - r + 1$. There are in total $n - 1$ possible r values and n different $A_{s_i,i}$ values, resulting in at most n^2 tests.

The $p_{r/n}^{\text{new}}$ values for all $r = 2, \ldots, 18$ are shown in Table 2b. Compared with Bonferroni BHPC p-values, the new GBHPC p-values can be much smaller especially when r is small. Both methods give a 95% confidence interval of the true proportion of non-null hypotheses r_0/n as $r_0/n \in [0.72, 1]$. At higher confidence levels the new method has a greater range for the number of non-null hypotheses, i.e., a smaller range of nulls are then compatible with the data.

6 Conclusion and future work

Partial conjunction hypotheses are natural hypotheses to test for measuring repeated effects across settings/studies. The null is to be rejected only when at least r hypotheses are non-null. By testing PC hypotheses at different r values, one can also construct a confidence interval of r_0/n, the true proportion of non-null hypotheses. Here a wider interval for non-null hypotheses indicates a narrower range for the number of nulls, that is, wider intervals are more conclusive.

This paper characterizes the admissible p-values for a partial conjunction test of independent hypotheses or hypotheses with positively associated P-values, within the class of non-decreasing p-values. Any monotone admissible p-value for $H_0^{r/n}$ is the maximum of the non-decreasing p-values for the global null in each combination of $n - r + 1$ hypotheses, which we call GBHPC p-values. We have shown that for sensitive GBHPC p-values, as long as each meta-analysis p-value of the

$n-r+1$ hypotheses is admissible, the combined p-value is monotone admissible. A consequence is that among combined p-values that only depend on the order statistics of individual p-values, the original BHPC p-values are the only monotone admissible ones. We also have found inadmissibility of GBHPC p-values without the monotonicity constraint. However, the dominated tests only have a moderate power gain at low power regions in the alternative space. Since these counter-examples are not monotone, they are hard to justify in practice and are not reasonable choices.

In summary, we illustrated the properties of tests for a PC hypothesis and characterized a class of good tests called GBHPC p-values. Compared with its symmetric form, the BHPC p-values, GBHPC p-values have more flexibility to adapt to complicated problem structure, and can therefore gain power at important regions in the alternative space, as we showed in our simulations and real data examples. The computational cost of non-symmetric GBHPC p-values can be of a concern, but there are special cases where GBHPC p-values are computable. One of the future directions is to expand the applications where computable GBHPC p-values are available.

Finally, there are variations of partial conjunctions that are useful in practice. For example, the count of replicability may vary for different hypotheses. Replication of effects from two distinct classes can be of more interest than replication in two similar classes. Another variation is to require that a null hypothesis is rejected only when there are at least r non-nulls with the same sign of effect. Such hypotheses can have very complex alternative and null space, and it can be the future work to understand their properties.

References

Benjamini, Y. and R. Heller (2008). Screening for partial conjunction hypotheses. *Biometrics 64*(4), 1215–1222.

Benjamini, Y., R. Heller, and D. Yekutieli (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 367*(1906), 4255–4271.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics 29*(4), 1165–1188.

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association 49*(267), 559–574.

Birnbaum, A. (1955). Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *The Annals of Mathematical Statistics 26*(1), 21–36.

Brown, L. D. and J. I. Marden (1989). Complete class results for hypothesis testing problems with simple null hypotheses. *The Annals of Statistics 17*(1), 209–235.

Esary, J. D., F. Proschan, and D. W. Walkup (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics 38*(5), 1466–1474.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics 9*(5), e1003486.

Friston, K. J., A. P. Holmes, C. J. Price, C. Büchel, and K. J. Worsley (1999). Multisubject fMRI studies and conjunction analyses. *Neuroimage 10*(4), 385–396.

Friston, K. J., W. D. Penny, and D. E. Glaser (2005). Conjunction revisited. *NeuroImage 25*(3), 661–667.

Heller, R. and D. Yekutieli (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics 8*(1), 481–498.

Higgins, J., S. G. Thompson, and D. J. Spiegelhalter (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 137–159.

Holland, P. W. and P. R. Rosenbaum (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics 14*(4), 1523–1543.

Iwasa, M. (1991). Admissibility of unbiased tests for a composite hypothesis with a restricted alternative. *Annals of the Institute of Statistical Mathematics 43*(4), 657–665.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *The Annals of Mathematical Statistics 23*(4), 541–552.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses.* Springer Science & Business Media.

Marden, J. I. (1982). Combining independent noncentral chi squared or F tests. *The Annals of Statistics 10*(1), 266–277.

Matthes, T. K. and D. R. Truax (1967). Tests of composite hypotheses for the multivariate exponential family. *The Annals of Mathematical Statistics 38*(3), 681–697.

Nichols, T., M. Brett, J. Andersson, T. Wager, and J.-B. Poline (2005). Valid conjunction inference with the minimum statistic. *Neuroimage 25*(3), 653–660.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics 37*(6B), 3867–3892.

Perlman, M. D. and L. Wu (1999). The emperor's new tests. *Statistical Science 14*(4), 355–369.

Price, C. J. and K. J. Friston (1997). Cognitive conjunction: A new approach to brain activation experiments. *NeuroImage 5*(4), 261–270.

Ruff, C. T., R. P. Giugliano, E. Braunwald, E. B. Hoffman, N. Deenadayalu, M. D. Ezekowitz, A. J. Camm, J. I. Weitz, B. S. Lewis, A. Parkhomenko, T. Yamashita, and E. M. Antman (2014). Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. *The Lancet 383*(9921), 955–962.

Sarkar, T. K. (1969). Some lower bounds of reliability. Technical report, DTIC Document.

Shenhav, L., R. Heller, and Y. Benjamini (2015). Quantifying replicability in systematic reviews: the r-value. Technical Report arxiv1502.00088, Tel-Aviv University.

Stein, C. (1956). The admissibility of Hotelling's T^2-test. *The Annals of Mathematical Statistics 27*(3), 616–623.

Wang, W., Z. Wei, and W. Sun (2010). Simultaneous set-wise testing under dependence, with applications to genome-wide association studies. *Statistics and Its Interface 3*, 501–511.

Zaykin, D. V., L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir (2002). Truncated product method for combining p-values. *Genetic epidemiology 22*(2), 170–185.

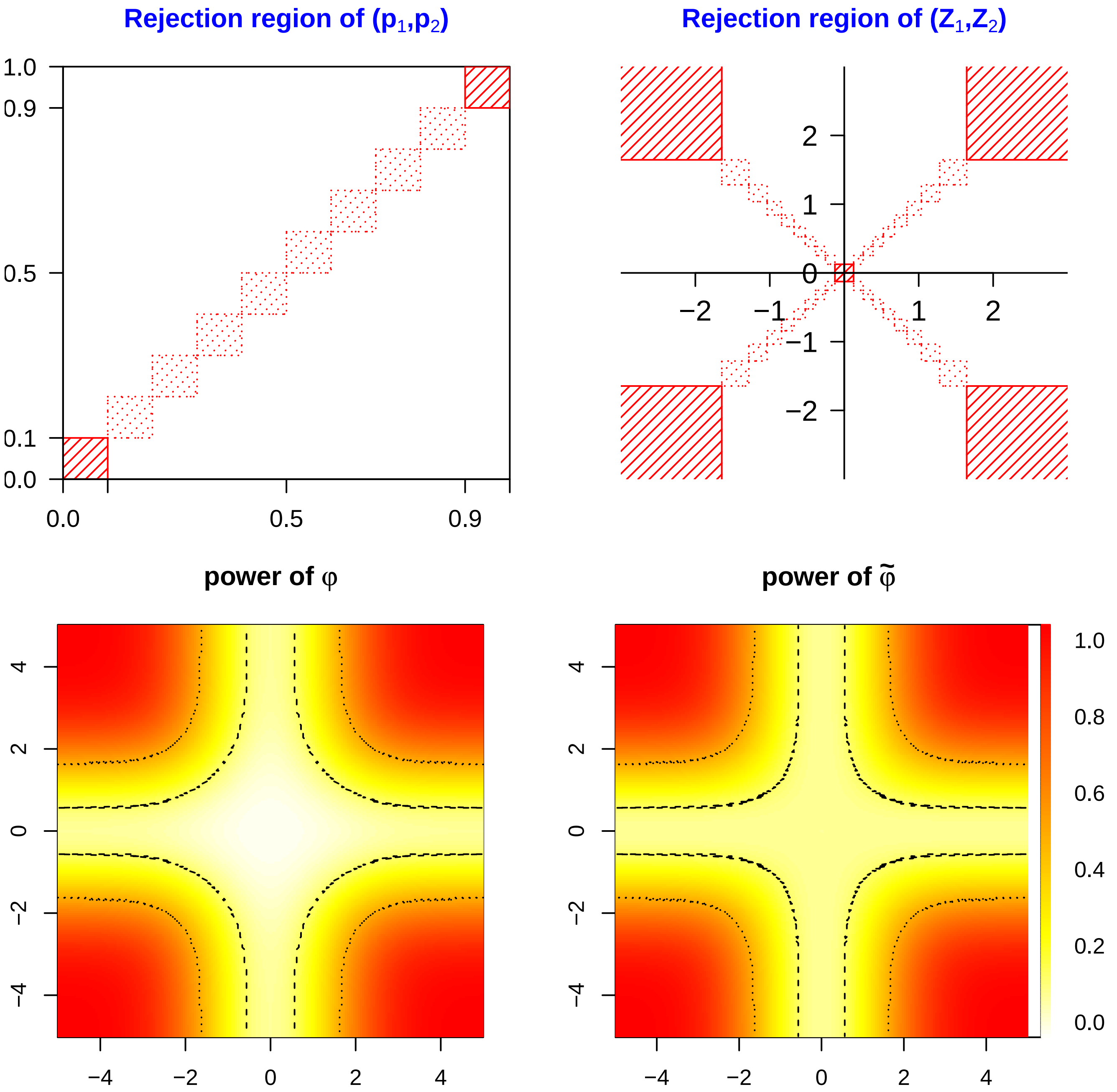

Figure 1: The top two plots: rejection regions of φ' and $\widetilde{\varphi}$ in the p-value space and the test statistic space, using $\alpha = 0.2$. The sold shaded region is the rejection region of φ', while the rejection region of $\widetilde{\varphi}$ also includes in the dotted shaded squares. The Bottom two plots: power comparison of φ and $\widetilde{\varphi}$. The dashed line is where power is at 0.15 and the dotted line is where the power is 0.5.

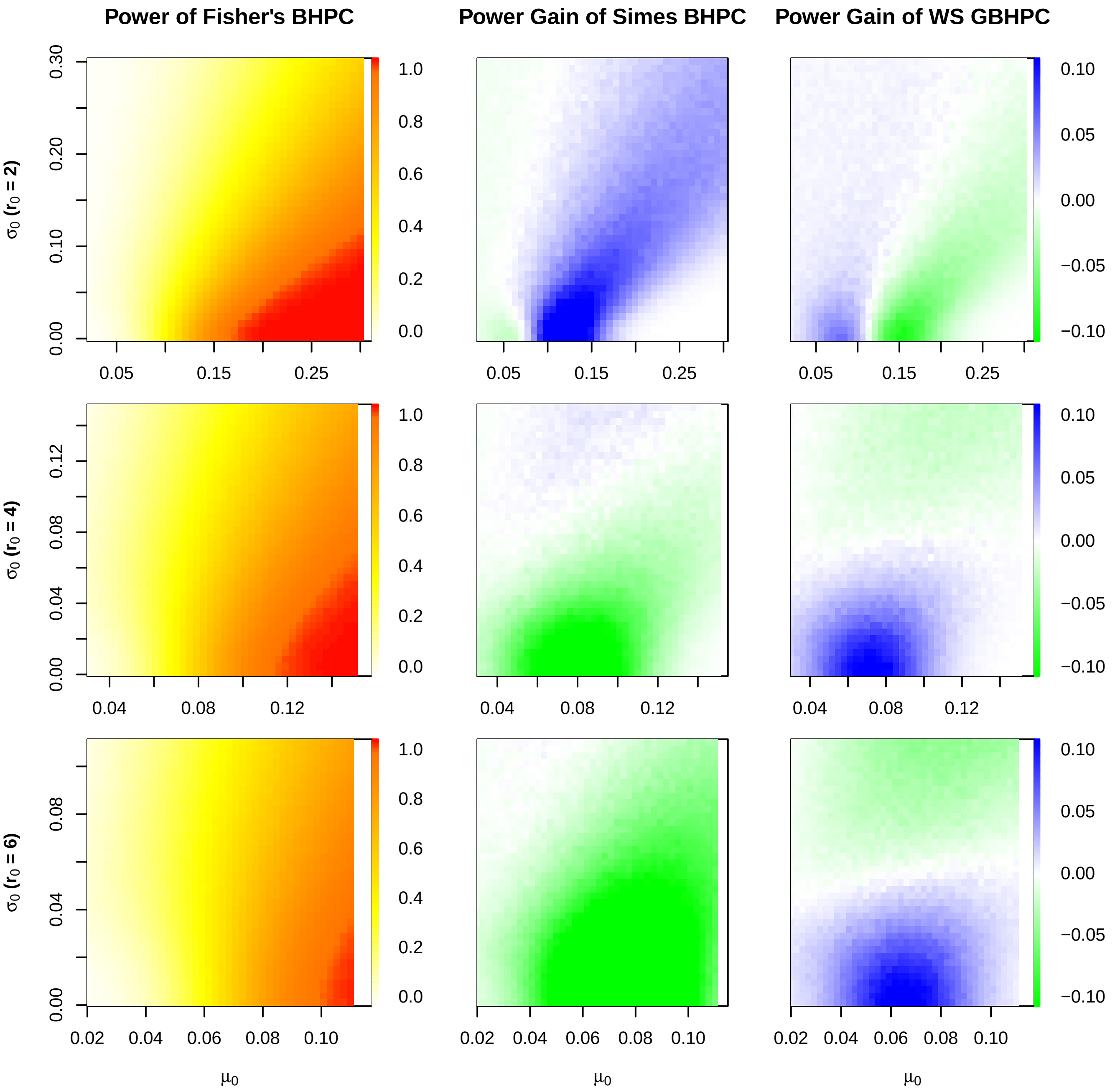

Figure 2: Power comparison of GBHPC p-values: Each row is for one r_0 value and each column is for one form of GBHPC p-value. the significance level is $\alpha = 0.05$. The first column are power maps Fisher's BHPC p-value against (μ_0, σ_0). the last two columns are the power difference of Simes and Weighted Stouffer's GBHPC p-values against Fisher's p-value. The green color indicates a power loss compared with Fisher's BHPC p-value while the blue color shows a power gain. All plots have a resolution of 40 × 40.

(a)

	Pooled NOAC (events)	Pooled Warfarin (events)	Estimated Odds Ratio	p value
Age(years)				
< 75	496/18073	578/18004	0.85	9.26E-03
≥ 75	415/11188	532/11095	0.76	6.61E-05
Sex				
Female	382/10941	478/10839	0.78	5.00E-04
Male	531/18371	634/18390	0.83	2.38E-03
Diabetes				
No	622/20216	755/20238	0.82	2.93E-04
Yes	287/9096	356/8990	0.79	3.81E-03
Previous stroke or TIA				
No	483/20699	615/20637	0.78	4.65E-05
Yes	428/8663	495/8635	0.85	2.14E-02
Creatinine clearance (mL/min)				
< 50	249/5539	311/5503	0.79	6.24E-03
50–80	405/13055	546/13155	0.74	5.85E-06
> 80	256/10626	255/10533	1.00	9.64E-01
CHADS$_2$ score				
0–1	69/5058	90/4942	0.75	7.83E-02
2	247/9563	290/9757	0.87	1.05E-01
3–6	596/14690	733/14528	0.80	5.21E-05
VKA status				
Naive	386/13789	513/13834	0.75	2.19E-05
Experienced	522/15514	597/15395	0.86	1.61E-02
Centre-based TTR				
$< 66\%$	509/16219	653/16297	0.78	2.49E-05
$\geq 66\%$	313/12742	392/12904	0.80	4.68E-03

(b)

r	$p^{\text{Bon}}_{r/n}$	$p^{\text{new}}_{r/n}$
2	3.73E-04	**4.49E-05**
3	3.98E-04	**4.66E-05**
4	6.98E-04	**7.50E-05**
5	7.29E-04	**1.18E-04**
6	8.59E-04	**1.31E-04**
7	3.52E-03	**1.39E-04**
8	5.50E-03	**4.23E-04**
9	2.38E-02	**1.90E-02**
10	3.43E-02	**2.66E-02**
11	3.75E-02	**2.81E-02**
12	4.37E-02	**4.63E-02**
13	**5.56E-02**	6.45E-02
14	8.07E-02	**6.45E-02**
15	8.56E-02	**7.36E-02**
16	2.35E-01	**7.36E-02**
17	2.11E-01	2.11E-01
18	9.64E-01	9.64E-01

(c)

	Age	**Sex**	**Diabetes**	**Stroke or TIA**	**Creatinine**	**CHADS$_2$**	**VKA**	**TTR**
$A_{1,i}$	9.26E-03	2.38E-03	3.81E-03	2.14E-02	9.64E-01	1.05E-01	1.61E-02	4.68E-03
$A_{2,i}$	9.37E-06	1.74E-05	1.64E-05	1.47E-05	3.68E-02	4.78E-02	5.61E-06	1.98E-06
$A_{3,i}$	–	–	–	–	5.83E-06	5.29E-05	–	–

Table 2: (a) The original data and individual p-values: the blue color highlights individual p-values that are not significant at level $\alpha = 0.05$; (b) the values of Bonferroni BHPC p-value and the new GBHPC p-values when r changes in $H_0^{r/n}$; (c) the grouping factor level combined p-values $A_{s,i}$ as defined in (7).