Multiple testing of partial conjunction null hypotheses with conditional \( p \)-values based on combination test statistics

Thorsten Dickhaus†

Institute for Statistics, University of Bremen, Bremen, Germany.
E-mail: dickhaus@uni-bremen.de

Ruth Heller

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv-Yafo, Israel.
E-mail: ruheller@tauex.tau.ac.il

Anh-Tuan Hoang

Institute for Statistics, University of Bremen, Bremen, Germany.
E-mail: anhtuan.hoang@uni-bremen.de

Summary. The partial conjunction null hypothesis is tested in order to discover a signal that is present in multiple studies. We propose methods for multiple testing of partial conjunction null hypotheses which make use of conditional \( p \)-values based on combination test statistics. Specific examples comprise the Fisher combination function and the Stouffer combination function. The conditional validity of the corresponding \( p \)-values is proved for certain classes of one-parametric statistical models, including one-parameter natural exponential families. The standard approach of carrying out a multiple test procedure on the (unconditional) partial conjunction \( p \)-values can be extremely conservative. We suggest alleviating this conservativeness, by eliminating many of the conservative partial conjunction \( p \)-values prior to the application of a multiple test procedure. This leads to the following two step procedure: first, select the set with partial conjunction \( p \)-values below a selection threshold; second, within the selected set only, apply a family-wise error rate or false discovery rate controlling procedure on the conditional partial conjunction \( p \)-values. By means of computer simulations and real data analyses, we compare the proposed methodology with other recent approaches.

Keywords: Data-adaptive multiple test, false discovery rate, family-wise error rate, Schweder-Spjøtvoll estimator, simultaneous statistical inference, uniform conditional stochastic order

†Address for correspondence: Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bibliothekstr. 5, D-28359 Bremen, Germany.
1. Introduction

"Replicability is widely taken to ground the epistemic authority of science." (Romero (2019)). In fact, the replication of scientific results is essential for their acceptance by the scientific community. In order to assess whether a scientific result has indeed been replicated in an independent study, appropriate scientific methods are needed. During recent years, statistical methods have been developed in this context; see, e.g., Bogomolov and Heller (2013), Heller et al. (2014), Heller and Yekutieli (2014), Bogomolov and Heller (2018), Wang and Owen (2019), Hung and Fithian (2020), and Hoang and Dickhaus (2021c). In the aforementioned articles, it has been proposed to formalize the replication of a certain finding as a statistical hypothesis which can be tested on the basis of a data sample by employing an appropriate statistical test procedure. Especially in the context of modern high-throughput technologies, the simultaneous testing of many (say m >> 1) non-replicability null hypotheses (corresponding to m different features or endpoints, respectively) is of considerable interest, connecting the replicability assessment with the theory of multiple hypothesis testing.

The no-replicability null hypothesis is a specific instance of a partial conjunction (PC) null hypothesis (Benjamini and Heller, 2008), defined as follows: given s individual null hypotheses and γ ∈ {2, ..., s}, at most γ − 1 individual null hypotheses are false. Thus, a rejected PC null hypothesis leads to the conclusion that at least γ individual null hypotheses are false. In replicability analysis, this means that the result is replicated in at least γ studies. PC null hypotheses are also used for other types of inference: for example, in Benjamini and Heller (2008), in order to identify the brain voxels in which at least γ out of s covariates are associated with the outcome; in Sun and Wei (2011), in order to identify the genes expressed in at least γ out of s time points; in Karmakar and Small (2020), in order to discover the outcomes with at least γ out of s evidence factors; in Li et al. (2021), in order to identify the genetic segments containing distinct association with the phenotype in at least γ out of s diverse environments.

Regarded as a subset of the parameter space of a statistical model, a PC null hypothesis is a composite null hypothesis, such that standard methods for computing a corresponding p-value can be very conservative; cf. Dickhaus (2013). One approach to overcome this conservativity is to exploit concepts from selective inference (see, among others, Fithian et al. (2017) and Zhao et al. (2019)).

In the present work, we elaborate on selective inference methods for multiple testing of PC null hypotheses. A two-stage multiple test procedure is proposed which first selects promising features (or endpoints) by means of their (conventional) p-values arising from a combination test for replicability. The p-values of the so-selected features get adjusted for the selection event by conditioning on the latter. In the second stage of testing, the conditional p-values are used in a (standard) multiple test. A key mathematical result will be that the proposed conditional p-values are valid (in the sense of Equation (1) in Hoang and Dickhaus (2021c)). This property will imply type I error control of the proposed two-stage multiple test. We will illustrate these theoretical points with prototypical statistical models and by
analyzing simulated as well as real data.

Wang et al. (2021) provide another interesting approach to address the conservativeness of PC $p$-values. They use a clever filtration and the Bonferroni combining method for testing multiple PC hypotheses. We compare and contrast our method with theirs, as well as with the direct approach of applying a multiple test procedure on the PC $p$-values, while highlighting the potential advantages of our conditional approach.

The remainder of the work is structured as follows. In Sections 2 and 3 we present our proposed statistical methodology. Section 2 describes the proposed conditional $p$-value for one single PC null hypothesis and discusses its uniform validity. Section 3 explains how a family of such conditional $p$-values can be used for multiple testing of a family of PC null hypotheses. Section 4 is devoted to exemplary statistical models to which our considerations apply. In Section 5, computer simulations are presented, and Section 6 deals with an application in the context of genome-wide association studies. We conclude with a discussion in Section 7. A variety of additional results is presented in the supplementary material (SM).

2. Proposed conditional $p$-value and its validity

Given a set of null hypotheses $H_1, \ldots, H_s$ together with their corresponding stochastically independent random $p$-values $P_1, \ldots, P_s$, and given a constant $1 \leq \gamma \leq s$, we are interested in testing the partial conjunction null hypothesis

$$H^\gamma/s = \{\text{at most } \gamma - 1 \text{ null hypotheses are false}\}.$$ 

We assume that the individual $p$-values are valid, i.e., $\Pr(P_i \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ if $H_i$ is true. We also assume that a combination $p$-value $P^\gamma/s(P_1, \ldots, P_s)$, that is increasing in $P_i, i = 1, \ldots, s$, is valid for $H^\gamma/s$ (see examples in Section 4). If $H^\gamma/s$ is true, the least favorable parameter configuration (LFC) is the one that maximizes the probability of rejection, $\Pr(P^\gamma/s \leq \alpha)$.

The LFC typically leads to $\gamma - 1$ $p$-values that are zero almost surely, and the remaining $s - \gamma + 1$ $p$-values are uniform. We assume that $P^\gamma/s$ is uniform under any LFC $\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})$ in $H^\gamma/s$, where $\pi$ is any permutation vector of $s$ elements and $U_1, \ldots, U_{s-\gamma+1}$ are stochastically independent and identically $\text{Uni}[0,1]$-distributed, where $\text{Uni}[0,1]$ denotes the (continuous) uniform distribution on the interval $[0,1]$. Among those that fulfill these assumptions are for example the Fisher, Stouffer and Simes combination functions. For more information about PC null hypotheses and suitable combination $p$-values $P^\gamma/s$ we refer to Benjamini and Heller (2008) and Hoang and Dickhaus (2021a).

Our goal is to find conditions for $P^\gamma/s$ and $P_1, \ldots, P_s$ such that, for a $\tau \in (0, 1)$, the conditional $p$-value $P^\gamma/s/\tau$ given $P^\gamma/s \leq \tau$ is also valid for $H^\gamma/s$. This turns out to be very useful when we consider a family of PC null hypotheses in the following sections, since we expect that far less than $\tau$ of the true PC null hypotheses will have PC $p$-values at most $\tau$, due to their conservativeness. Thus by selecting all PC $p$-values at most $\tau$, we greatly reduce the multiplicity problem. But in order to use the conditional PC $p$-values on the reduced family of selected PC hypotheses, we
need to prove they are indeed valid (given selection). We define sufficient conditions for validity in Section 2.1 and prove that the conditional PC p-values are indeed valid in Section 2.2.

Throughout, we use the following notation. For any random variable $X$ and event $A$, let $[X|A]$ denote any random variable whose distribution is the conditional distribution of $X$ given $A$. Furthermore, we let $\leq_{st}$ denote the usual stochastic order, $\leq_{rh}$ the reversed hazard rate (rh) order, $\leq_{lr}$ the hazard rate (hr) order, and $\leq_{lr}$ the likelihood ratio (lr) order, cf. Sections 1.A.–1.C. in Shaked and Shanthikumar (2007).

2.1. The set-up
We denote by $\theta = (\theta_1, \ldots, \theta_s)^\top$ the parameter (vector) of the statistical model under consideration for all data ascertained in all $s$ studies. For each $i \in \{1, \ldots, s\}$, the (marginal) test problem is $H_i : \theta_i \geq \theta_i^*$ versus $K_i : \theta_i < \theta_i^*$, where $(\theta_i^* : 1 \leq i \leq s)$ are given constants. We assume that the distribution of $P_i$ is independent of $\theta_{i'}$, for all $i' \neq i$. For each $i \in \{1, \ldots, s\}$ we assume the $p$-value to be valid, i.e., $P_{\theta_i}(P_i \leq \alpha) \leq \alpha$ for all $\alpha \in [0,1]$ and $\theta_i \geq \theta_i^*$. Furthermore, we require that it holds
\[
P_{\theta_i}(P_i \leq \alpha) = \alpha
\]
for all $\alpha$, i.e., that $P_i$ is uniformly distributed under the parameter $\theta_i^*$. This requirement can be removed, but it simplifies the exposition (it typically holds for continuous test statistics). We address the more general setting without requirement (1) in Section 4 following the example of multiple binomial tests. We need $P_i$ to satisfy conditional validity as well, i.e., $P_{\theta_i}(P_i/\tau \leq \alpha \mid P_i \leq \tau) \leq \alpha$ for all $\alpha \in [0,1]$ and $\theta_i \geq \theta_i^*$. This is equivalent to requiring that (Zhao et al., 2019)
\[
\forall t \leq \tau, \theta_i \geq \theta_i^*: P_{\theta_i}(P_i \leq t \mid P_i \leq \tau) \leq \frac{t}{\tau}.
\]

Let $P_{\theta_i}^{(\theta_i)}$ denote a random variable with the same distribution as $P_i$, when the data is generated from the distribution indexed by the parameter $\theta_i$. The usual validity can also be written as $\text{Uni}[0,1] \leq_{st} P_{\theta_i}^{(\theta_i)}$ for all $\theta_i \in H_i$, and the equation (2) holding for all $\tau \in [0,1]$, i.e. conditional validity for all $\tau$, is equivalent to $\text{Uni}[0,1] \leq_{lr} P_{\theta_i}^{(\theta_i)}$ for all $\theta_i \in H_i$, which is stronger.

Thus, we assume that one of the following two conditions hold:

(A1) It holds that $\text{Uni}[0,1] \leq_{rh} P_{\theta_i}^{(\theta_i)}$, for all $\theta_i \in H_i$, $i = 1, \ldots, s$.

(A2) For all $\theta_i, \bar{\theta}_i \in \mathbb{R}$ with $\theta_i \leq \bar{\theta}_i$, it holds that $P_{\theta_i}^{(\theta_i)} \leq_{rh} P_{\theta_i}^{(\bar{\theta}_i)}$, $i = 1, \ldots, s$.

Condition (A2) is stronger than (A1). For testing multiple PC null hypotheses, assuming (A1) is enough. We provide the result in Section 2.2 assuming (A2) since it may be of independent interest, see Remark 1.

A known result is that the likelihood ratio order implies the reversed hazard rate order, cf. Theorem 1.C.1. in Shaked and Shanthikumar (2007). Thus, $p$-values
that are isotone or antitone transformations of test statistics that are likelihood ratio ordered in the parameter \( \theta \), fulfill condition (A2). The latter property is also frequently referred to as monotone likelihood ratio (MLR) of the considered test statistics. For example any one-parametric, linear exponential family fulfills this MLR property with respect to its sufficient statistic, cf. Karlin and Rubin (1956b).

2.2. Main result

The conditional \( p \)-value \( P^{\gamma/s}/\tau \) given \( P^{\gamma/s} \leq \tau \) is valid for \( H^{\gamma/s} \) if and only if

\[
\frac{\mathbb{P}_\theta(P^{\gamma/s} \leq t)}{\mathbb{P}_\theta(P^{\gamma/s} \leq \tau)} \leq \frac{t}{\tau},
\]

for all \( t \in [0, \tau] \), and \( \theta \in H^{\gamma/s} \), see for example Zhao et al. (2019).

The condition in (3) is equivalent to \([U \mid U < \tau] \leq_{st} \{P^{\gamma/s} \mid P^{\gamma/s} < \tau\}(\theta)\), where \( U \) is a \([0, 1]\)-distributed random variable. Therefore, if \( U \leq_{rh}(P^{\gamma/s}(\theta)) \), we get (3) for all \( \tau \), and therefore validity of the conditional \( p \)-value for all \( \tau \) (cf. Shaked and Shanthikumar 2007, Section 1.B.6). We call \( P^{\gamma/s} \) uniformly valid, if its conditional \( p \)-value is valid for all \( \tau \).

**Theorem 1 (Uniform conditional validity).**

(i) If (A1) holds, then \( \text{Univ}[0, 1] \leq_{rh} (P^{\gamma/s}(\theta)) \), for all \( \theta \in H^{\gamma/s} \).

(ii) If (A2) holds, then \( (P^{\gamma/s}(\theta)) \leq_{rh} (P^{\gamma/s}(\tilde{\theta})) \), for all \( \theta, \tilde{\theta} \in \mathbb{R}^s \), where \( \theta \leq \tilde{\theta} \), component-wise.

In particular, (A1) and (A2) imply that \( P^{\gamma/s}(P_1, \ldots, P_s) \) is a uniformly valid \( p \)-value for \( H^{\gamma/s} \). Furthermore, (A2) implies that the distributions \(((P^{\gamma/s}(\theta)))_\theta\) are rh-ordered with respect to every component of \( \theta \).

**Proof.** We show the second statement first. To this end, let \( \theta, \tilde{\theta} \in \mathbb{R}^s \), where only the first components \( \theta_1, \tilde{\theta}_1 \) of the two vectors are different, with \( \theta_1 \leq \tilde{\theta}_1 \).

Under Condition (A2), it holds \( P_1(\theta) \leq_{rh} P_1(\tilde{\theta}) \). Since \( p_1 \mapsto P^{\gamma/s}(p_1, p_2, \ldots, p_s) \) is non-decreasing for all given values \( p_2, \ldots, p_s \), it follows

\[
P^{\gamma/s}(P_1, p_2, \ldots, p_s)(\theta) \leq_{rh} P^{\gamma/s}(P_1, p_2, \ldots, p_s)(\tilde{\theta})
\]

from Theorem 1.B.43 in Shaked and Shanthikumar (2007) for all given \( p_2, \ldots, p_s \). Since we assumed that \( P_1 \) and \( P_2 \) are stochastically independent, \([P^{\gamma/s}(P_1, p_2, \ldots, p_s) \mid P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq \tau]\) and \([P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau]\) have the same distribution, for any given \( p_2, \ldots, p_s \), and any \( \tau \). Therefore,

\[
\mathbb{P}_\theta(P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau) = \int F_{\theta|p_2, \ldots, p_s, \tau}(t) d\mathbb{P}_{\theta|p_2}(p_2),
\]

where \( F_{\theta|p_2, \ldots, p_s, \tau}(t) = \mathbb{P}_\theta(P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq \tau) \) and \( \mathbb{P}_{\theta|p_2} \) is the distribution of \( P_2 \) under \( \theta \). With analogous notations for \( \tilde{\theta} \), we have that
\( \mathbb{P}_\theta = \mathbb{P}_{\tilde{\theta}} \), because \( \theta, \tilde{\theta} \) do not differ in their second components and \( P_2 \) does not depend on the other components \( \theta_i, i' \neq 2 \).

For any fixed \( \tau \), it holds \( F_{\theta|P_2, \ldots, P_s, \tau}(t) \geq F_{\tilde{\theta}|P_2, \ldots, P_s, \tau}(t) \) from (4), so that (5) and \( \mathbb{P}_{\theta} = \mathbb{P}_{\tilde{\theta}} \) imply

\[
\mathbb{P}_\theta(P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau) \\
\geq \mathbb{P}_{\tilde{\theta}}(P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau)
\]

for all \( p_3, \ldots, p_s \) and any \( \tau \). But this means

\[
P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \mathbb{P}_\theta P^{\gamma/s}(P_1, P_2, \ldots, p_s) \leq \tau.
\]

Analogously, since \( P_1, P_2, P_3 \) are jointly stochastically independent,

\[
[P^{\gamma/s}(P_1, P_2, p_3, p_4, \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, p_3, p_4, \ldots, p_s) \leq \tau]
\]

and

\[
[P^{\gamma/s}(P_1, P_2, p_3, p_4, \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, p_3, p_4, \ldots, p_s) \leq \tau] \mid P_3 = p_3
\]

have the same distribution. We can (iteratively) argue as before, and at the end, we have

\[
P^{\gamma/s}(P_1, \ldots, P_s) \leq \mathbb{P}_\theta P^{\gamma/s}(P_1, \ldots, P_s) \leq \tau.
\]

The proof for any \( \theta, \tilde{\theta} \in \mathbb{R}^s \), that only differ in their \( i \)-th components (instead of their first ones) is analogous. Since the relation \( \leq \) is transitive, we obtain the statement (ii) of the theorem. Namely, in the case of arbitrary parameter vectors \( \theta \leq \tilde{\theta} \), we can apply the above reasoning successively to each coordinate \( i \) in which \( \theta \) and \( \tilde{\theta} \) differ.

For establishing statement (i) of the theorem, we assume (A1) instead of (A2). We assumed that, under an LFC of \( P^{\gamma/s} \) under \( H^{\gamma/s} \),

\[
P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}))
\]

is uniformly distributed, where \( \pi \) is any permutation vector of \( s \) elements. For any \( \theta \in H^{\gamma/s} \), there are at least \( s - \gamma + 1 \) components \( \theta_i \) with \( \theta_i \in H_i \), and thus for at least \( s - \gamma + 1 \) indices \( i \), it holds \( \text{Uni}[0,1] \leq \mathbb{P}(P_i(\theta)). \) For the remaining \( \gamma - 1 \) marginal \( p \)-values it necessarily holds \( 0 \leq \mathbb{P}(\theta). \) Thus, there exists a permutation (vector) \( \pi \) such that \( \pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}) \) is in the rh order component-wise smaller than \( (P_1, \ldots, P_s) \). As in the proof of statement (ii), one can then show

\[
P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})) \leq \mathbb{P}(P_1, \ldots, P_s) \leq \tau \]

by sequentially replacing each component on the left-hand side by the (in rh order) larger \( p \)-value on the right-hand side. This means \( \text{Uni}[0,1] \leq \mathbb{P}(P_1, \ldots, P_s) \) and therefore uniform validity of \( P^{\gamma/s}(P_1, \ldots, P_s) \) for \( H^{\gamma/s} \).
Remark 1. Theorem 1 is closely related to the notion of the uniform conditional stochastic order (UCSO). This notion has been discussed, among others, by Whitt (1980, 1982), and by Lynch et al. (1987).

Remark 2. Instead of requiring that $P^{\gamma/s}$ is uniform under any LFC of the form $\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})$ in $H^{\gamma/s}$, where $\pi$ is any permutation vector of $s$ elements, the condition

$$\text{Uni}[0,1] \leq m P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}))$$

would have been sufficient for Theorem 1. For an example of a combination function $P^{\gamma/s}$ that does not fulfill the above condition, we refer to Section S7 of the SM.

3. Testing multiple partial conjunction null hypotheses

In this section, we assume that $m > 1$ (marginal) PC null hypotheses $H_1^{\gamma/s}, \ldots, H_m^{\gamma/s}$ are simultaneously under consideration under the scope of one and the same statistical model. In this, the assumption that $\gamma$ and $s$ are the same for all $1 \leq j \leq m$ is not necessary, and it is only made for notational convenience.

If all marginal test problems $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$, $1 \leq j \leq m$, are such that our Theorem 1 applies to each of them, $m$ valid conditional (or randomized) $p$-values can readily be obtained. Many standard multiple test procedures like, for instance, the famous linear step-up test by Benjamini and Hochberg (1995) (the BH procedure) for control of the false discovery rate (FDR), have as their main assumption regarding the marginal $p$-values on which they operate that the latter marginal $p$-values are valid. (Oftentimes, some further conditions regarding the dependency structure among the marginal $p$-values have to be fulfilled, but transforming unconditional $p$-values into conditional $p$-values for each $j \in \{1, \ldots, m\}$ separately does not alter the latter dependency structure.)

On the basis of the aforementioned considerations, we propose for multiple testing of PC hypotheses the following workflow.

Algorithm 3.1.

(i) Compute for each $j \in \{1, \ldots, m\}$ an unconditional $p$-value $P_j^{\gamma/s}$ for testing $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$.

(ii) Choose cutoffs $\tau_1, \ldots, \tau_m$. For all those coordinates $j$ for which $P_j^{\gamma/s} > \tau_j$, retain $H_j^{\gamma/s}$.

(iii) In the case that there exists at least one index $j$ with $P_j^{\gamma/s} \leq \tau_j$: Transform, for each $j \in \{1, \ldots, m\}$ with $P_j^{\gamma/s} \leq \tau_j$ separately, the marginal unconditional $p$-value $P_j^{\gamma/s}$ into a marginal conditional $p$-value as described in Section 2. For continuous test statistics, the resulting conditional PC $p$-value is $P_j^{\gamma/s}/\tau_j$. 
(iv) Utilize the conditional $p$-values obtained in step (iii) in a standard multiple test procedure $\varphi$ (say), which merely requires validity of the marginal $p$-values on which it operates.

For additional power enhancement, we propose to employ a data-adaptive multiple test procedure in step (iv) which makes use of a pre-estimate of the proportion $\pi_0$ of true null hypotheses among the selected ones and incorporates the estimate in its decision rule. Adaptive procedures are especially useful when the fraction of null hypothesis is small. Even if in the family of $m$ PC null hypotheses considered, the fraction of PC null hypotheses is close to one, following selection the fraction of true PC null hypotheses among the selected may be far smaller than one.

For the sake of simplicity, we focus on the case $\tau_1 = \cdots = \tau_m = \tau$ and provide a brief summary of approaches for selecting $\tau$. Let $S_\tau$ be the set of those indices in $\{1, \ldots, m\}$ for which the unconditional $p$-values are not greater than $\tau$.

1. **(No conditioning)** With $\tau = 1$ we use the unconditional $p$-values $P_{1}^{\gamma/s}, \ldots, P_{m}^{\gamma/s}$ in step (iv).

2. **(Pre-specified $\tau$)** Choose a $\tau \in (0, 1)$ beforehand. Multiple testing in step (iv) is done on $\{p_{j}/\tau \mid j \in S_\tau\}$.

3. **(Adaptive choice of $\tau$)** Proposition 4 in Zhao et al. (2019) provides an adaptive way, based on the $p$-values $p_{1}^{\gamma/s}, \ldots, p_{m}^{\gamma/s}$, of choosing $\tau$ that retains the validity of any valid global test, if $P_{1}^{\gamma/s}, \ldots, P_{m}^{\gamma/s}$ are jointly stochastically independent and uniformly valid. Given a sequence $0 \leq \tau_1 < \cdots < \tau_K \leq 1$ of $\tau$’s between 0 and 1, we go from $\tau_k$ to $\tau_{k-1}$, starting with $\tau_K$, if the $p$-values $\{p_{j}^{\gamma/s} \mid j \notin S_{\tau_k}\}$ greater than $\tau_k$ fulfill certain conditions. The idea is to only use $\{p_{j}^{\gamma/s} \mid j \notin S_{\tau}\}$ to (adaptively) choose $\tau$, and to use $\{p_{j}^{\gamma/s}/\tau \mid j \in S_{\tau}\}$ in steps (ii) – (iv). Zhao et al. (2019) give an example of a condition on the $p$-values $\{p_{j}^{\gamma/s} \mid j \notin S_{\tau}\}$ motivated by the Bonferroni test. The (Bonferroni-) adjusted $p$-values $p_{1}^{\gamma/s}/|S_{\tau}|/\tau, j = 1, \ldots, m$, are minimized as functions of $\tau$, if $|S_{\tau}|/\tau$ is minimized. For more details, we refer to Section 3.3 in Zhao et al. (2019). Since $|S_{\tau}| = m\hat{F}_{m}(\tau)$, where $\hat{F}_{m}$ is the empirical cumulative distribution function (ecdf) of the $p$-values $P_{1}^{\gamma/s}, \ldots, P_{m}^{\gamma/s}$, adaptive minimization of any function of $G(\tau) = H(|S_{\tau}|, \tau)$, where $H$ is increasing in $|S_{\tau}|$ and decreasing in $\tau$, as in Zhao et al. (2019), would retain Proposition 4 in Zhao et al. (2019). In Section S5 in the SM we give another example of one such approach.

### 3.1. Theoretical properties of multiple test procedures targeted for FDR control

Applying the BH procedure on $\{p_{j}^{\gamma/s} \mid j \in S_{\tau}\}$, for a fixed pre-specified $\tau$, guarantees control of the FDR for the entire family of PC hypotheses if all $m \times s$ $p$-values are stochastically independent, under the conditions of Theorem 1. This clearly follows since the BH procedure controls the FDR on the set of valid and independent $p$-values (Benjamini and Hochberg, 1995), and $\{p_{j}^{\gamma/s} \mid j \in S_{\tau}\}$ is such a set. Using
the same reasoning, the adaptive BH procedure on \( \{ p_{j}^{\gamma/s}, j \in S_{\tau} \} \) also guarantees control of the FDR if we use Storey’s estimator (Storey et al., 2004) for the fraction of null hypotheses in \( S_{\tau} \). In this section we provide theoretical guarantees for FDR control when the within-study \( p \)-values are dependent (which is the norm in high-dimensional studies), or when \( \tau \) is chosen adaptively.

First, for the BH procedure using a pre-specified \( \tau \), we show that the FDR is controlled at the nominal level if the original within study \( p \)-values are independent or positive regression dependent on the subset (PRDS) of true null hypotheses (Benjamini and Yekutieli, 2001).

**Proposition 3.1.** Assume that the conditions of Theorem 1 are satisfied for each \( p \)-value. Moreover, assume that for each study, the \( p \)-values are PRDS on the subset of \( p \)-values corresponding to true null hypotheses, and the \( p \)-values across studies are independent. In addition, assume that the PRDS property is preserved for every subset of \( p \)-values under marginalization over the remaining \( p \)-values. Then the FDR of the BH procedure at level \( \alpha \) on \( \{ p_{j}^{\gamma/s} / \tau, j \in S_{\tau} \} \) for a fixed pre-specified \( \tau \), is at most \( \alpha \).

**Proof.** According to Theorem 4.1 in Bogomolov (2021), the level \( \alpha \) BH procedure on the PC \( p \)-values, \( \{ p_{j}^{\gamma/s}, j = 1, \ldots, m \} \), guarantees that the FDR on the PC null hypotheses is controlled at level \( \alpha \) if the \( p \)-values are PRDS on the subset of \( p \)-values corresponding to true null hypotheses within each study, and the \( p \)-values across studies are independent. Since the dependence structure of any subset of \( p \)-values is unchanged, any subset taken is also PRDS within each study on the subset of \( p \)-values corresponding to true null hypotheses. In addition, Theorem 1 guarantees for the subset indexed by \( S_{\tau} \), that \( \{ p_{j}^{\gamma/s} / \tau, j \in S_{\tau} \} \) are valid \( p \)-values. Thus the level \( \alpha \) BH procedure on \( \{ p_{j}^{\gamma/s} / \tau, j \in S_{\tau} \} \) controls the FDR for the family of null hypotheses \( \{ H_{j}^{\gamma/s}, j \in S_{\tau} \} \), as well as unconditionally for \( \{ H_{j}^{\gamma/s}, j = 1, \ldots, m \} \).

The following proposition is a slight generalization of Proposition 4 of Zhao et al. (2019) and deals with the case of an adaptively chosen \( \tau \) in the context of Algorithm 3.1.

**Proposition 3.2.** Let \( \theta \in \Theta \) be arbitrary, but fixed. Assume that \( (P_{j}^{\gamma/s}, 1 \leq j \leq m) \) are jointly stochastically independent. Assume that the multiple test \( \varphi \) employed in step (iv) of Algorithm 3.1 is such that

\[
\mathbb{E}_{\theta} \left[ g(V_{m}, R_{m}) \right] \leq \alpha \tag{7}
\]

holds true for any fixed value of \( \tau \in (0, 1] \), where \( V_{m} \) is the (random) number of type I errors of \( \varphi \), \( R_{m} \) the (random) total number of rejections of \( \varphi \), \( g \) some measurable function taking values in \([0, 1]\), and \( \alpha \in (0, 1) \) some given constant.

†This is satisfied, e.g., under the subset pivotality condition 2.1 in Westfall and Young (1993), or when the dependence structure of every subset of \( p \)-values is preserved under marginalization over the remaining \( p \)-values.
Now, let $F_x = \sigma\left(\{P_j^{\gamma/s} : P_j^{\gamma/s} \geq x\}\right)$ for $x \in [0,1]$, and assume that $\tilde{\tau}$ is a backward stopping time in the sense that the event $\{\tilde{\tau} \geq x\}$ is $F_x$-measurable for any $x \in [0,1]$. Then, (7) remains true if the fixed value of $\tau$ is replaced by the random value of $\tilde{\tau}$.

**Proof.** Proposition 4 of Zhao et al. (2019) yields the assertion for the special case of $g(V_m, R_m) = \mathbb{I}\{V_m > 0\}$ and for $\theta$ in the global null hypothesis. (Actually, the test $\varphi$ considered by Zhao et al. (2019) is just a single test for the global null hypothesis $H_0$ (say), such that $\mathbb{P}_\theta(V_m > 0)$ reduces to the type I error probability of that single test under $\theta \in H_0$.) However, as already indicated by Zhao et al. (2019) in their Section 3.4, the proof of their Proposition 4, which is presented in their Appendix A.3, does neither make use of the specific form of the function $g$ nor of the fact that $\theta \in H_0$ is assumed. Therefore, their proof applies to general (measurable and integrable) functions $g$ and to arbitrary parameter values $\theta \in \Theta$.

The next proposition provides an asymptotic FDR control guarantee for the following greedy choice of $\tau$: the value which leads to the largest number of rejections in step (iv) of Algorithm 3.1 when using the level $\alpha$ BH procedure. This value cannot be too small (since in this case too many false PC null hypotheses are not selected) nor too large (since in this case too many stochastically larger than uniform PC $p$-values are among the selected). The choice of this greedy $\tau$ can be written concisely as follows. Recall that the BH cutoff for any family of null hypotheses of size $K$ is $x_{BH} = \max\{x : \frac{K \times x}{R(x)} \leq \alpha\}$, where $R(x)$ is the number of $p$-values at most $x$, and all hypotheses with $p$-values at most $x_{BH}$ are rejected. In step (iv) of Algorithm 3.1, the number of hypotheses is $K = |S_{\tau}|$, and the number of conditional PC $p$-values at most $x$ equals the number of unconditional PC $p$-values at most $\tau x$, i.e., $|S_{\tau x}|$. More formally, for any $x \in [0,1]$,

$$R(x) = \sum_{i \in S_{\tau}} \mathbb{I}\left(\frac{p_i^{\gamma/s}}{\tau} \leq x\right) = \sum_{i \in S_{\tau}} \mathbb{I}\left(p_i^{\gamma/s} \leq \tau x\right) = \sum_{i=1}^m \mathbb{I}\left(p_i^{\gamma/s} \leq \tau x\right) = |S_{\tau x}|.$$

Denoting the BH threshold for a given $\tau$ by $\hat{x}(\tau)$:

$$\hat{x}(\tau) = \max \left\{x : \frac{|S_{\tau}| \times x}{|S_{\tau x}| \vee 1} \leq \alpha \right\},$$

we choose $\hat{\tau} = \arg \max_{\tau \in \{\tau_1, \ldots, \tau_K\}} S_{\tau \hat{x}(\tau)}$, where $0 < \tau_1 < \ldots < \tau_K \leq 1$ is a pre-defined finite set of $K$ candidate values for the selection threshold. The false discovery proportion (FDP) of the rejections made with $(\hat{\tau}, \hat{x}(\hat{\tau}))$ is asymptotically almost surely (a.s.) at most $\alpha$, assuming the following limits exist:

$$\forall x \in (0,1] : \lim_{m \to \infty} \frac{V_m(x)}{m_0} = G_0(x) \text{ and } \lim_{m \to \infty} \frac{R_m(x) - V_m(x)}{m - m_0} = G_1(x) \text{ a.s.,} \quad (8)$$

where $m_0$ is the number of true PC null hypotheses; $V_m(x)$ is the (random) number of true PC null hypotheses with $p$-values below $x$; $R_m(x)$ is the number of (random)
p-values below $x$; and $G_0$ and $G_1$ are continuous functions such that

$$\forall x \in (0, 1]: 0 < G_0(x) \leq x; \quad (9)$$

$$\lim_{m \to \infty} \frac{m_0}{m} = \pi_0 \text{ exists.} \quad (10)$$

**Proposition 3.3.** Assume that condition (A1) or (A2) is satisfied for each $p$-value. Moreover, assume that the convergence assumptions of equations (8)-(10) hold for the PC $p$-values $\{p_j^{\gamma/s}, j = 1, \ldots, m\}$. Then, the FDP of the BH procedure at level $\alpha$ on $\{p_j^{\gamma/s}/\hat{\tau}, j \in S_\hat{\tau}\}$ is asymptotically at most $\alpha$.

See Section S3 in the SM for the proof.

**4. Illustrative example applications**

In this section, we exemplify applications of our proposed methodology in the context of two widely used statistical model classes. Namely, we consider multiple $Z$-tests in Gaussian shift models and multiple binomial tests for success parameters of Bernoulli distributions.

**Model 1 (Multiple $Z$-tests).** For given sample sizes $n_{i,j}$, assume that we can observe $\{X_k^{(i,j)}: i = 1, \ldots, s, j = 1, \ldots, m, k = 1, \ldots, n_{i,j}\}$, and that for each study $i$ and coordinate $j$ the observables $X_1^{(i,j)}, \ldots, X_{n_{i,j}}^{(i,j)}$ are stochastically independent and identically normally distributed on $\mathbb{R}$ with expected value $\theta_{i,j}$ and a known variance, which may without loss of generality be assumed to be equal to one. For the study- and endpoint-specific test problem $H_{i,j} = \{\theta_{i,j} \geq \theta^*_{i,j}\}$ versus $K_{i,j} = \{\theta_{i,j} < \theta^*_{i,j}\}$, we consider the test statistic $T_{i,j} = n_{i,j}^{-1/2} \sum_{k=1}^{n_{i,j}} (X_k^{(i,j)} - \theta^*_{i,j})$, which is normally distributed on $\mathbb{R}$ with expected value $\sqrt{n_{i,j}} \cdot (\theta_{i,j} - \theta^*_{i,j})$ and variance one, for all $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, m\}$. The corresponding $p$-variable is given by $P_{i,j} = \Phi (T_{i,j})$, where $\Phi$ denotes the cumulative distribution function (cdf) of the standard normal distribution on $\mathbb{R}$.

**Remark 3.** If the variance of $X_1^{(i,j)}$ is unknown under Model 1, we can instead consider the $t$-distributed Studentized means as test statistics $T_{i,j}$. Then (A2) is still fulfilled, because the family of non-central $t$-distributions $\left(\frac{T_{i,j}^{(\theta_{i,j})}}{\theta_{i,j}}\right)$ is likelihood ratio ordered with respect to the non-centrality parameter, cf. Karlin and Rubin (1956a).

**Model 2 (Multiple binomial tests).** For given sample sizes $n_{i,j}$, assume that we can observe $\{X_k^{(i,j)}: i = 1, \ldots, s, j = 1, \ldots, m, k = 1, \ldots, n_{i,j}\}$, and that for each study $i$ and coordinate $j$ the observables $X_1^{(i,j)}, \ldots, X_{n_{i,j}}^{(i,j)}$ are stochastically independent and identically Bernoulli-distributed indicator variables with success parameter $\pi_{i,j}$, which we assume to lie in the open interval $(0, 1)$, to avoid...
pathologies. The corresponding canonical parameter of the resulting linear exponential family is given by \( \theta_{i,j} = \log(\pi_{i,j}) = \log(\pi_{i,j}/(1 - \pi_{i,j})) \). The study- and endpoint-specific test problem \( H_{i,j} = \{ \theta_{i,j} \geq \theta^*_j \} \) versus \( K_{i,j} = \{ \theta_{i,j} < \theta^*_j \} \) is related to the original parameter values by noticing that \( \pi_{i,j} \mapsto \theta_{i,j} = \log(\pi_{i,j}) \) is a strictly increasing (thus one-to-one) transformation of \( \pi_{i,j} \in (0,1) \) onto \( \mathbb{R} \). We denote the value of the success parameter corresponding to the value \( \theta^*_j \) of the canonical parameter by \( \pi^*_i,j \). Moreover, we consider the test statistic \( T_{i,j} = \sum_{k=1}^{n_{i,j}} X_k^{(i,j)} \), which is binomially distributed on \( \{0, \ldots, n_{i,j}\} \) with parameters \( n_{i,j} \) and \( \pi_{i,j} \) for all \( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, m\} \). The corresponding (random) \( p \)-value is given by \( P_{i,j} = F_{\text{Bin}(n_{i,j}, \pi^*_i,j)}(T_{i,j}) \), where \( F_{\text{Bin}(n_{i,j}, \pi^*_i,j)} \) denotes the cdf of the binomial distribution with parameters \( n_{i,j} \) and \( \pi^*_i,j \).

Under Model 2, the base \( p \)-values \( P_{i,j} \) are discrete and do not fulfill our assumptions from Section 2. More particularly, one can show that discrete \( p \)-values with finite support can never be greater than \( \text{Uni}[0,1] \) in \( \text{rh} \) order, and can therefore never be uniformly valid (unless the \( p \)-value is a.s. equal to one). Therefore, we introduce a more general definition of conditional \( p \)-values that coincide with the version from Section 2 if the \( p \)-value is continuously distributed. As before, we consider the discrete \( p \)-value \( P^{\gamma/s}_{j}(P_{1,j}, \ldots, P_{s,j}) \), \( j = 1, \ldots, m \).

**Definition 1.** The conditional \( p \)-values we consider in Model 2, and more generally in discrete models, are \( \left( P^{\gamma/s}_{j}\{P^{\gamma/s}_{j} \leq \tau \} \mid P^{\gamma/s}_{j} \leq \tau \right)_{j=1,\ldots,m} \), where \( \theta^*_j \) is an LFC parameter for \( H^{\gamma/s}_{j} \).

Note that it holds \( \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) < \tau \), if \( \tau \) is not in the support of \( P^{\gamma/s}_{j} \), but it a.s. holds \( P^{\gamma/s}_{j} \leq P_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \) if \( P^{\gamma/s}_{j} \leq \tau \). If \( P^{\gamma/s}_{j} \) is uniformly distributed under LFCs, this definition coincides with the one from Section 2.

**Lemma 1.** If it holds \( P^{\gamma/s}_{j}(\theta^*_j) \leq \text{rh} \ (P^{\gamma/s}_{j}(\theta^*_j) \text{ for all } \theta^*_j \in H^{\gamma/s}_{j} \), then \( P^{\gamma/s}_{j}/P_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \) given \( P^{\gamma/s}_{j} \leq \tau \) is valid for all \( \tau \in (0,1] \).

**Proof.** For \( \theta^*_j \in H^{\gamma/s}_{j} \), we have to show that it holds

\[
\mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq t \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right) \leq t \mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right), \quad t \in [0,1].
\]

We assumed that \( \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq t) \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \) is increasing in \( t \), thus, it holds

\[
\frac{\mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq t \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right)}{\mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right)} \leq \frac{\mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right)}{\mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right)}.
\]

Note, that the cdf of \( P^{\gamma/s}_{j} \) under \( \theta^*_j \) is the identity function on the set of the support points, and thus it holds \( \mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right) = \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \). With \( \mathbb{P}_{\theta^*_j} \left( P^{\gamma/s}_{j} \leq t \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \right) \leq t \mathbb{P}_{\theta^*_j}(P^{\gamma/s}_{j} \leq \tau) \), the proof is complete.
That Model 2 fulfills the condition in Lemma 1 follows from the fact that the set of Bernoulli distributions is monotone likelihood ratio ordered with respect to the success parameter $\theta_{i,j}$, for all $i, j$. The rest is similar to Part (i) in Theorem 1, which does not require continuously distributed $p$-values.

Under Models 1 and 2, we may consider the following combination $p$-values.

(i) Fisher combination: For each $j \in \{1, \ldots, m\}$, we let

$$P_j^{\gamma/s} = 1 - F_{\chi^2_{2(s-\gamma+1)}} \left( -2 \sum_{i=\gamma}^{s} \log (P_{i:s,j}) \right),$$

where the notation $P_{i:s,j}$ refers to order statistics, and $F_{\chi^2_{2(s-\gamma+1)}}$ denotes the cdf of the chi-square distribution with $2(s - \gamma + 1)$ degrees of freedom; cf. Section 21.1 in Fisher (1934).

(ii) Stouffer combination: For each $j \in \{1, \ldots, m\}$, we let

$$P_j^{\gamma/s} = 1 - \Phi \left( \frac{1}{\sqrt{s-\gamma+1}} \sum_{i=\gamma}^{s} \Phi^{-1} (1 - P_{i:s,j}) \right);$$

cf. Footnote 14 in Section V of Chapter 4 of Stouffer et al. (1949).

By making use of the aforementioned properties of the family of normal or Student’s $t$ distributions on $\mathbb{R}$ and of the family of Bernoulli distributions with a fixed number of trials and success parameter in $(0, 1)$, respectively, it is straightforward to check that the assumptions of Theorem 1 are fulfilled under Models 1 and 2, both for the Fisher combination and for the Stouffer combination.

**Remark 4.** Under Model 1, choosing the Stouffer combination method in a fixed coordinate $j$ leads to the so-called ”Gaussian selection differential” as the test statistic for testing $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$; cf., e. g., Nagaraja (1981) and Section 6.3 in Tong (1990). Hence, Theorem 1 implies the reverse hazard rate order for the selection differential in a Gaussian shift model with respect to shifts in the mean.

5. Computer simulations

To assess and compare the performance of different replicability analysis approaches, we have carried out simulation studies. We have four specific aims: (1) to compare the power of our conditional approach with the unconditional approach, as well as with the “indirect” approach, adaFilter (Wang et al., 2021); (2) to study the effect of dependency within each study; (3) to examine how the choice of selection threshold matters, as well as the performance of an adaptive selection of this threshold for the conditional approach; (4) to examine the use of adaptive procedures in our conditional approach, which make use of an estimate of the fraction of null hypotheses on the selected hypotheses.
We consider the following data generation setting, which is partly inspired by the Crohn’s disease example considered in Section 6, where we analyse \( s = 8 \) GWAS studies, and it is estimated that about 90% of the SNPs have no negative association with the disease in all the studies. We set the number of PC hypotheses based on \( s = 8 \) independent studies to \( m = 20000 \), \( \left\{ H_{i}^{\gamma}, i = 1, \ldots, 20000 \right\} \), for \( \gamma \in \{2, 3, 4, 5\} \). The fraction of true global null hypotheses (i.e., for which all \( s \) elementary hypotheses are true) is 0.9; the fraction of false null PC hypotheses is \( \pi_1 \in \{0.002, 0.02\} \), where each non-null PC hypothesis has equal probability; each of the remaining true null PC hypotheses also has equal probability, adding up to an additional \( 1 - 0.9 - \pi_1 \) true null PC hypotheses (but false global nulls). This type of data generation was considered in Wang et al. (2021) and made available in their function \( \text{GenPMat()} \) in their R package implementing AdaFilter available at https://github.com/jingshuw/adaFilter. We use their package both for generating data from the Gaussian shift model and for the adaFilter analysis (described in Section 1). The test statistics are generated from the Gaussian shift model, with non-null means sampled independently from \( |\theta_{i,j}| \in \{3, 4\} \).

Within each study the test statistics are Gaussian, with a block correlation structure. The covariance within each block is symmetric with off diagonal entry \( \rho = 0.9 \) and diagonal entry one. The block size is 10 (see Section S4.2 in the SM for results with block size 100).

We apply the replicability analysis detailed in (i)-(iv) of Section 3 for a selection threshold \( \tau \in (0, 1) \). In step (i), the PC \( p \)-values are computed using the Fisher combination method (the use of Stouffer and Simes combination methods is evaluated in Section S4.1 in the SM).

In steps (ii)-(iii), \( \tau_j \) is either fixed for all hypotheses at 0.1, or chosen adaptively as in Zhao et al. (2019). The adaptive approach of choosing \( \tau \) by Zhao et al. (2019), as briefly described in Section 3, continues from \( \tau_k \) to \( \tau_{k-1} \) if we can reject \( q > \hat{F}_m(\tau_{k})/\tau_k \), where \( m(\hat{F}_m(\tau_{k}+\omega) - \hat{F}_m(\tau_{k})) \sim \text{Binomial}(m, q\omega) \) at a pre-specified significance level \( \beta \), and \( \omega \leq 1 - \tau_K \). Above, \( \hat{F}_m \) denotes the empirical cumulative distribution function of the \( p \)-values \( \{p_{\gamma/s}^{\gamma/s}, \ldots, p_{m}^{\gamma/s}\} \). Note that this stopping condition only uses the \( p \)-values \( \{p_{\gamma/s}^{\gamma/s}, j \notin S_{\tau_k} \} \) larger than \( \tau_k \). Since their method assumes the test statistics are independent, we apply it on a subsample of independent test statistics (taking into consideration only statistics that are approximately 1.5 times the block size apart). Moreover, we examine several values of their parameter \( \beta \): the larger the value of \( \beta \), the smaller the estimate of the adaptive threshold. The window size is \( w = 0.1 \) as in Zhao et al. (2019).

In step (iv), the following two multiple testing procedures are applied to the conditional \( p \)-values for FDR control at the 0.05 level: the Benjamini-Hochberg (BH, Benjamini and Hochberg 1995) procedure at level 0.05; the adaptive BH procedure suggested in Storey et al. (2004) which applies the BH procedure at level 0.05/\( \hat{\pi}_0 \), where \( \hat{\pi}_0 \) is (a slight variation on) the plug-in estimator of Schweder and Spjotvoll (1982) for the fraction of true PC null hypotheses following selection (i.e., among all hypotheses with PC \( p \)-value at most \( \tau \)).

We compare this analysis with the unconditional approach of applying the BH
procedure at level 0.05 on the PC \( p \)-values. We also compare to the approach of Wang et al. (2021): first filter the potential hypotheses using the PC \( p \)-values based on the Bonferroni combination method, for the null PC hypotheses \( \{H_i^{(\gamma-1)/8}, i = 1, \ldots, m\} \); then identify as discoveries the subset of filtered hypotheses for which the estimated FDP is at most 0.05, by using the PC \( p \)-values based on the Bonferroni combination method, for the null PC hypotheses of interest \( \{H_i^{\gamma/8}, i = 1, \ldots, m\} \).

Tables 1 and 2 show the average number of true discoveries (our notion of power) and FDP for the novel procedures, using the selection threshold pre-specified as \( \tau = 0.1 \) or adaptively chosen as in Zhao et al. (2019) with \( \beta = 0.1, 0.5 \), versus competitors. Compared with the unconditional approach (which corresponds to \( \tau = 1 \)) of applying BH on the PC \( p \)-values, we see that the power is greater with the novel approach, for every \( \gamma \). This is expected, since the threshold for discovery for each selected hypothesis is lower using the conditional approach. An intuitive explanation is as follows. Figure 1 shows the distribution of the number of hypotheses selected for each \( \gamma \). The number selected divided by the threshold for selection, \(|S_\tau|/\tau|\), is much smaller than \( m = 20000 \). Selection tends to eliminate many more PC null hypotheses than the expected number \( \tau \times (1 - \pi_1) \times m \), because most \( p \)-values have a null distribution that is stochastically much larger than uniform, see Section S1 in the SM for details. So had we considered doing the Bonferroni procedure following selection, each selected hypothesis would have been rejected if the PC \( p \)-value is at most \( \alpha \times \tau/|S_\tau| > \alpha/m \). Therefore, if all the nonnulls (with enough power for detection) are selected, the conditional approach has more power than the unconditional approach. This reasoning carries over also to BH instead of Bonferroni, as Table 1 shows. The power advantage over the unconditional approach ranges from a power increase of (at least) 6% for \( \gamma = 2 \) to a power increase of more than 200% for \( \gamma = 5 \). From Table 2 it is clear that the conditional and unconditional procedures are below the nominal 0.05 level, and that the unconditional approach is the most conservative (i.e., with lowest FDR level). This conservatism is due to the fact that most PC \( p \)-values have a null distribution that is stochastically much larger than uniform, and the conditional \( p \)-values have a null distribution that is closer to uniform, but still conservative, see Section S1 in the SM for details.

Compared with adaFilter, we see in Table 1 that the power is greater with the novel approach when \( \pi_1 = 0.002 \) but not when \( \pi_1 = 0.02 \). The setting with \( \pi_1 = 0.02 \) is more favorable to adaFilter (over \( \pi_1 = 0.002 \)), since the ratio of false \( \gamma/s \) PC hypotheses to false \( (\gamma - 1)/s \) PC hypotheses is larger (due to the fact that we keep the number of true global null hypotheses unchanged as \( \pi_1 \) increases), so the selection step of adaFilter is more efficient. In Table 2 we see that adaFilter controls the FDR but is less conservative than the other procedures.

Since the fraction of true PC hypotheses is close to one, the adaptive BH procedure on the PC \( p \)-values does not have a power advantage over BH so we do not compare this method. We see in Table 1 that when \( \pi_1 = 0.002 \), the adaptive approach also does not have a power advantage over BH on the conditional PC \( p \)-values, but when \( \pi_1 = 0.02 \), the adaptive approach makes more discoveries on average. This is so because the advantage of adaptivity increases as \(|S_\tau|\) contains
Table 1. In the symmetric block dependent setting, the average number of true discoveries for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC $p$-values, BH and adaptive BH (denoted aBH) on conditional PC $p$-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

| $\pi_1$ | $\gamma$ | adaFilter | BH   | $\hat{\tau} = 0.1$ | BH   | aBH | $\hat{\tau}$ with $\beta = 0.1$ | BH   | aBH | $\hat{\tau}$ with $\beta = 0.5$ | BH   | aBH |
|---------|----------|-----------|------|-------------------|------|-----|--------------------------------|------|-----|--------------------------------|------|-----|
| 0.002   | 2        | 25.3      | 25.7 | 28.4              | 28.2 | 28.4 | 28.3                           | 28.6 | 28.6|
|         | 3        | 17.0      | 14.5 | 20.1              | 20.1 | 19.8 | 19.6                           | 20.1 | 20.3|
|         | 4        | 10.5      | 6.2  | 12.0              | 12.2 | 11.5 | 11.2                           | 11.8 | 11.8|
|         | 5        | 7.0       | 2.1  | 6.4               | 6.6  | 5.9  | 5.7                            | 6.2  | 6.1 |
| 0.02    | 2        | 338.5     | 302.4| 319.9             | 327.9| 320.5| 321.8                          | 320.4| 325.5|
|         | 3        | 281.3     | 202.3| 242.7             | 268.7| 249.7| 254.2                          | 249.5| 259.8|
|         | 4        | 217.8     | 110.1| 160.1             | 195.9| 176.1| 182.7                          | 176.1| 189.2|
|         | 5        | 168.7     | 46.8 | 91.7              | 124.0| 112.8| 121.0                          | 112.7| 127.0|

For example, for $\gamma = 2$, Figure 1 shows that hundreds of hypotheses are selected, so when $\pi_1 = 0.002$ the fraction of PC nulls is very close to one (since only $20000 \times 0.002 = 40$ PC hypotheses are nonnull) but when $\pi_1 = 0.02$ the fraction of PC nulls can be far from one (since $20000 \times 0.02 = 400$ PC hypotheses are nonnull and most of them are likely to be among the selected). The advantage of the adaptive procedure is greater as $\gamma$ increases: for $\gamma = 2$ only a smaller fraction of nulls. For example, for $\gamma = 2$, Figure 1 shows that hundreds of hypotheses are selected, so when $\pi_1 = 0.002$ the fraction of PC nulls is very close to one (since only $20000 \times 0.002 = 40$ PC hypotheses are nonnull) but when $\pi_1 = 0.02$ the fraction of PC nulls can be far from one (since $20000 \times 0.02 = 400$ PC hypotheses are nonnull and most of them are likely to be among the selected). The advantage of the adaptive procedure is greater as $\gamma$ increases: for $\gamma = 2$ only few additional discoveries are made on average, but for $\gamma = 5$ it is 10% or more.

Finally, we see from Table 1 that all choices of $\tau$ provide similar power for $\pi_1 = 0.002$, as well as for $\pi_1 = 0.02$ when $\gamma = 2$. However, for $\pi_1 = 0.02$ and $\gamma > 2$, the BH procedure on the pre-specified $\tau = 0.1$ has lower power compared with BH on the adaptively selected $\tau$. Moreover, the adaptive choice with $\beta = 0.5$ is at least as powerful as with $\beta = 0.1$ in all settings considered. Table 3 shows the average value of the adaptive selection threshold over 5000 repetitions for each simulations setting.

6. An application to Crohn’s disease genome-wide association studies

Identifying genomic regions with replicated association with Crohn’s disease is important for better understanding the disease pathogenesis. Franke et al. (2010) carried out a meta-analysis of eight genome-wide association (GWA) studies of Crohn’s disease in order to identify loci that are associated with the disease in at least one study. In this section, we illustrate our suggested replicability analysis in order to identify the SNPs associated with Crohn’s disease in at least $\gamma \in \{2, 3, 4, 5\}$ out of the eight studies.

For every study, the data consists of $z$ test statistics for the null hypothesis of no association between SNP and Crohn’s disease, for $m = 953,241$ autosomal SNPs. For each SNP, replication of association across studies can be defined with or without regard to the direction of association. For illustration purposes, we consider here only one direction. So our starting point for the replicability analysis...
Table 2. In the symmetric block dependent setting, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC $p$-values, BH and adaptive BH (denoted aBH) on conditional PC $p$-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

| $\pi_1$ | $\gamma$ | adaFilter | BH | BH $\tau = 0.1$ | BH $\tau$ with $\beta = 0.1$ | BH $\tau$ with $\beta = 0.5$ | BH aBH | aBH $\tau = 0.1$ | aBH $\tau$ with $\beta = 0.1$ | aBH $\tau$ with $\beta = 0.5$ |
|---------|--------|---------|-----|-----------------|-----------------|----------------|-------|-----------------|-----------------|----------------|
| 0.002   | 2      | 0.046   | 0.004 | 0.016 0.014     | 0.015 0.015     | 0.017 0.018    |       |                 |                 |
|         | 3      | 0.046   | 0.002 | 0.017 0.018     | 0.016 0.015     | 0.017 0.019    |       |                 |                 |
|         | 4      | 0.045   | 0.001 | 0.015 0.017     | 0.013 0.011     | 0.014 0.015    |       |                 |                 |
|         | 5      | 0.044   | 0.000 | 0.009 0.011     | 0.007 0.006     | 0.009 0.008    |       |                 |                 |
| 0.02    | 2      | 0.035   | 0.004 | 0.010 0.017     | 0.011 0.012     | 0.011 0.015    |       |                 |                 |
|         | 3      | 0.032   | 0.002 | 0.008 0.019     | 0.010 0.012     | 0.010 0.014    |       |                 |                 |
|         | 4      | 0.027   | 0.001 | 0.005 0.016     | 0.009 0.011     | 0.009 0.013    |       |                 |                 |
|         | 5      | 0.022   | 0.000 | 0.003 0.009     | 0.007 0.008     | 0.007 0.010    |       |                 |                 |

Table 3. In the symmetric block dependent setting, the average adaptively selected $\tau$ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

| $\pi_1$ | $\gamma$ | $\hat{\tau}$ with $\beta = 0.1$ | $\hat{\tau}$ with $\beta = 0.5$ |
|---------|--------|-------------------------------|-------------------------------|
| 0.002   | 2      | 0.11                          | 0.06                          |
|         | 3      | 0.16                          | 0.09                          |
|         | 4      | 0.25                          | 0.16                          |
|         | 5      | 0.33                          | 0.22                          |
| 0.02    | 2      | 0.22                          | 0.14                          |
|         | 3      | 0.27                          | 0.21                          |
|         | 4      | 0.35                          | 0.29                          |
|         | 5      | 0.42                          | 0.36                          |
Fig. 1. The distribution of the number of hypotheses selected, $|S_\tau|$, for every $\gamma$, in the simulation with $\pi_1 = 0.002$ (left column) and $\pi_1 = 0.02$ (right column), for $\tau$ selected as follows: pre-specified at $\tau = 0.1$ (top row), for $\tau$ selected as in Zhao et al. (2019) with $\beta = 0.1$ (middle row), and with $\beta = 0.5$ (bottom row). Based on 5000 repetitions.
is a matrix of $953,241 \times 8$ left sided $p$-values.

For FDR control at the 0.05 level, we apply the BH procedure or the adaptive BH procedure on the conditional $p$-values. The PC $p$-values are computed using the Fisher combination method on the left sided $p$-values, and the threshold for selection is either fixed or estimated from the data using the method of Zhao et al. (2019). Since the SNPs are dependent, only one in (approximately) 100 SNPs is used for estimation of the data adaptive threshold (the lag chosen was 100 since for the eight studies, the autocorrelation graph at lag 100 was very small). For comparison, we also applied adaFilter (Wang et al., 2021), BH on the PC $p$-values, and adaptive BH on the PC $p$-values, at the 0.05 level. (The analysis using the right sided $p$-values is omitted, since it provided qualitatively similar results in terms of the relative performance of the various methods.)

Figure 2 shows the number of rejections by each method for each $\gamma$. Our novel approach makes more discoveries than the unconditional approach of applying the BH or adaptive BH procedure on the PC $p$-values. Moreover, for a wide range of selection thresholds, the novel approach makes more discoveries than adaFilter. In particular, more discoveries are made using the adaptive thresholds: for $\gamma = 2$ (the minimal replicability requirement) more than twice as many SNPs are discovered. The advantage of the conditional approach over adaFilter for $\gamma < 5$ is due to the fact that a large fraction of the hypotheses rejected with $H^{(\gamma-1)/s}$ cannot be rejected with $H^{\gamma/s}$, so following the filtering stage adaFilter still faces a large multiplicity problem but with a less efficient test statistic (which is a single order statistic, thus it does not pool the information across studies by summation).

In order for the conditional approach with adaptive BH to make more discoveries than with BH, the estimated fraction of null PC hypotheses has to be below one. This occurs only when the selection threshold is small enough. Table 4 shows the number of selected PC hypotheses and the estimated fraction of null PC hypotheses among the selected when the selection threshold is pre-specified at $\tau = 0.1$, as well as when it is adaptively chosen by the method in Zhao et al. (2019) for $\beta = 0.1, 0.5$.

7. Discussion

We present a powerful approach for testing multiple PC hypotheses: first select the promising candidates, which are the PC hypotheses with PC $p$-values at most a certain threshold $\tau$; then apply a valid multiple testing procedure on the conditional PC $p$-values within the selected set only. Results from simulations and data analysis highlight the potential usefulness of our approach for the discovery of consistent signals across multiple studies.

For high dimensional studies, the test statistics within each study are typically dependent. Moreover, FDR may be preferred over FWER for controlling false positive findings. We expect our approach with FDR control to be highly robust to dependencies within each study, for the following reasons. First, we have a theoretical guarantee that for PRDS dependency within each study the finite sample FDR is controlled, and for local dependency within each study the asymptotic FDR is controlled. Second, a vast amount of empirical evidence suggests that the BH
Fig. 2. For every $\gamma$, the number of rejections at level 0.05, as a function of the selection threshold, using the conditional $p$-values for BH (solid black curve) and adaptive BH (dotted blue curve), as well as for adaFilter (dash-dotted green line). The number of rejections using the adaptive threshold with $\beta = 0.1$ and $\beta = 0.5$ for BH and for adaptive BH are in red circles and red pluses, respectively (with window size $w = 0.05$ in the algorithm of Zhao et al. 2019). The values at selection threshold ‘1.0’ correspond to no selection, i.e., the BH and adaptive BH on the (unconditional) PC $p$-values. For adaptive BH, we only display results with estimated fraction of null PC hypotheses below one.
Table 4. In the Crohn’s disease dataset, the threshold for selection, number selected, and estimated fraction of null PC hypotheses among the selected for the following three methods for threshold selection: pre-specified at 0.1, and adaptively selected $\tau$ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

| $\gamma$ | threshold selection method | $\tau$ | $|S_\tau|$ | estimated fraction of null PC hypotheses |
|---------|-----------------------------|--------|-------------|------------------------------------------|
| 2       | pre-specified at $\tau = 0.1$ | 0.10   | 32810       | 1.03                                      |
|         | Zhao et al. (2019) with $\beta = 0.1$ | 0.15   | 52575       | 1.09                                      |
|         | Zhao et al. (2019) with $\beta = 0.5$ | 0.01   | 3936        | 0.66                                      |
| 3       | pre-specified at $\tau = 0.1$ | 0.10   | 8771        | 1.09                                      |
|         | Zhao et al. (2019) with $\beta = 0.1$ | 0.05   | 3972        | 0.85                                      |
|         | Zhao et al. (2019) with $\beta = 0.5$ | 0.04   | 3240        | 0.82                                      |
| 4       | pre-specified at $\tau = 0.1$ | 0.10   | 2767        | 1.03                                      |
|         | Zhao et al. (2019) with $\beta = 0.1$ | 0.07   | 1860        | 0.94                                      |
|         | Zhao et al. (2019) with $\beta = 0.5$ | 0.06   | 1594        | 0.91                                      |
| 5       | pre-specified at $\tau = 0.1$ | 0.10   | 1006        | 0.89                                      |
|         | Zhao et al. (2019) with $\beta = 0.1$ | 0.14   | 1570        | 1.09                                      |
|         | Zhao et al. (2019) with $\beta = 0.5$ | 0.08   | 799         | 0.84                                      |

procedure controls the FDR at the nominal level for most dependencies occurring in practice, and this robustness carries over to our novel approach which applies the BH procedure on conditional PC $p$-values. Third, the dependency among PC $p$-values is less severe than within individual studies. As $\gamma$ increases, the PC $p$-values are less dependent, since the overlap between the identity of studies combined to form the PC $p$-values is reduced (and the studies are independent), see Section S2 in the SM for details. The competitor AdaFilter (Wang et al., 2021) does not provide a finite sample FDR guarantee for any type of dependence among the test statistics, but it does provide an asymptotic guarantee under assumptions of weak dependence.

The choice of the selection threshold $\tau$ can have a large effect on the power to detect false PC null hypotheses. In our numerical experiments, we show that for a wide range of $\tau$ values our conditional approach leads to greater power than the unconditional approach. We also show that for an asymptotic FDR guarantee, when the dependence is local within each study, it is possible to greedily choose the value of $\tau$ that leads to the greatest number of rejections; and for a finite sample FDR guarantee when the PC $p$-values are independent, it is possible to use the approach in Zhao et al. (2019). The data adaptive choice of Zhao et al. (2019) works quite well in our numerical experiments. From Proposition 3.2 it follows that other data adaptive methods that only use the information $|S_\tau|$ for choosing $\tau$ from a series of decreasing cutoffs are equally valid. In Section S5 of the SM we consider another data adaptive choice, which aims to balance the benefit of having a reduced selection set with the harm in inflating each PC $p$-value by the factor $1/\tau$ (in order to have a valid conditional PC $p$-value). The two data adaptive choices do not dominate each other, but it may be that for specific applications one choice is better than the other.

When the fraction of PC null hypotheses among the selected is small, the adap-
tive BH provides more discoveries than the BH procedure on the conditional PC $p$-values. For simplicity, we used Storey’s plug-in method for estimating the fraction of PC null hypotheses among the selected, as it is widely popular. However, other methods can be considered. In particular, methods of estimation of $\pi_0$ have been suggested in Hoang and Dickhaus (2021c,b) in settings where the $p$-values are conservative. Storey’s plug-in method tends to overestimate the fraction of nulls in such settings. Hoang and Dickhaus (2021c,b) address this conservatism by suggesting to use randomized $p$-values. The mathematical conditions for validity of the randomized $p$-values (in the sense of Dickhaus 2013) are the same as for our proposed conditional $p$-values. Hence, the considerations of Hoang and Dickhaus (2021c,b) can directly be applied in our present context, too. In Section S4.3 we show the advantage of using the estimate of Hoang and Dickhaus (2021c,b) when the conditional PC $p$-values are conservative.

Our proposed workflow can be used with many $p$-value combining methods. Our results use the Fisher combining method, since it has excellent power properties for a wide range of signals (Benjamini and Heller, 2008; Hoang and Dickhaus, 2021a). We consider the Stouffer and Simes combining methods, which are also quite popular, in Section S4.1. However, in many applications, some of the null $p$-values may be very conservative, and then even the global null (i.e., when testing with $\gamma = 1$) $p$-values based on Fisher, Stouffer, and Simes combining methods are conservative, since they assume that the $p$-values to combine are uniformly distributed. This is the case when the $p$-values are discrete, or when the individual null hypotheses are composite. For example, in the Crohn’s disease GWA studies, when testing for negative association, if the association is positive then the corresponding $p$-value will have a null distribution that is stochastically larger than uniform. Zhao et al. (2019) suggested for this purposes a test of the global null hypothesis that builds upon any combining function as follows: first, select for combining only the individual $p$-values that are at most a specified threshold; next, compute the conditional $p$-values on the selected set; finally, compute the global null conditional test $p$-value using the valid conditional $p$-values only. They demonstrated the power advantage of their conditional test over the unconditional global test (using, e.g., Fisher combining) when some $p$-values are conservative. The advantage is due to the fact that the unconditional test assumes that all $p$-values are uniformly distributed, but the conditional test only assumes this for the selected set. Thus, we expect our workflow in applications such as the Crohn’s disease GWA studies to provide even more discoveries using a combining method adapted to conservative $p$-values as suggested by Zhao et al. (2019). In Section S6 we demonstrate that when the original $p$-values are conservative, using the combining method of Zhao et al. (2019) leads to conditional PC $p$-values that have a more uniform null distribution, compared with the Fisher combining method considered in this manuscript. Combining their method and ours yields even better results, especially if the proportion of true null hypotheses is high. We leave for future work the comprehensive examination of the benefits of using such state-of-the-art combining functions with our proposed methodology.
Conditional combination test p-values

References

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

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol., 57, 289–300.

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

Bogomolov, M. (2021) Testing partial conjunction hypotheses under dependency, with applications to meta-analysis. arXiv:2105.09032.

Bogomolov, M. and Heller, R. (2013) Discovering findings that replicate from a primary study of high dimension to a follow-up study. J. Amer. Statist. Assoc., 108, 1480–1492. URL: https://doi.org/10.1080/01621459.2013.829002.

— (2018) Assessing replicability of findings across two studies of multiple features. Biometrika, 105, 505–516. URL: https://doi.org/10.1093/biomet/asx029.

Dickhaus, T. (2013) Randomized p-values for multiple testing of composite null hypotheses. J. Stat. Plann. Inference, 143, 1968–1979.

Fisher, R. A. (1934) Statistical methods for research workers. Fifth Edition. Oliver and Boyd, Edinburgh and London.

Fithian, W., Sun, D. and Taylor, J. (2017) Optimal inference after model selection. Preprint, available at https://arxiv.org/abs/1410.2597.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R. et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed croln’s disease susceptibility loci. Nature Genetics, 42, 1118–1125.

Heller, R., Bogomolov, M. and Benjamini, Y. (2014) Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. Proc. Natl. Acad. Sci. U.S.A., 111, 16262–16267.

Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. Ann. Appl. Stat., 8, 481–498. URL: https://doi.org/10.1214/13-AOAS697.

Hoang, A.-T. and Dickhaus, T. (2021a) Combining independent p-values in replicability analysis: A comparative study. arXiv preprint arXiv:2104.13081.

— (2021b) On the usage of randomized p-values in the Schweder-Spjøtvoll estimator. Forthcoming in the Annals of the Institute of Statistical Mathematics.
(2021c) Randomized p-values for multiple testing and their application in replicability analysis. *Biometrical Journal, early view.*

Hung, K. and Fithian, W. (2020) Statistical methods for replicability assessment. *The Annals of Applied Statistics, 14,* 1063 – 1087. URL: https://doi.org/10.1214/20-AOAS1336.

Karlin, S. and Rubin, H. (1956a) Distributions possessing a monotone likelihood ratio. *Journal of the American Statistical Association, 51,* 637–643.
— (1956b) The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics,* 272–299.

Karmakar, B. and Small, D. (2020) Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. *Annals of Statistics,* 48.

Li, S., Sesia, M., Romano, Y., Candes, E. and Sabatti, C. (2021) Searching for consistent associations with a multi-environment knockoff filter. arXiv:2106.04118v1.

Lynch, J., Mimmack, G. and Proshchan, F. (1987) Uniform stochastic orderings and total positivity. *The Canadian Journal of Statistics,* 15, 63–69.

Nagaraja, H. N. (1981) Some finite sample results for the selection differential. *Annals of the Institute of Statistical Mathematics,* 33, 437–448.

Romero, F. (2019) Philosophy of science and the replicability crisis. *Philosophy Compass, 14,* e12633. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12633.

Schweder, P. and Spjotvoll, E. (1982) Plots of p-values to evaluate many tests simultaneously. *Biometrika,* 69, 493–502.

Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic orders.* Springer Series in Statistics. Springer, New York. URL: https://doi.org/10.1007/978-0-387-34675-5.

Storey, J., Taylor, J. and Siegmund, D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B,* 66, 187–205.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams Jr, R. M. (1949) *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), vol. 1.* Princeton Univ. Press.

Sun, W. and Wei, Z. (2011) Multiple Testing for Pattern Identification, With Applications to Microarray Time-Course Experiments. *Journal of the American Statistical Association,* 106, 73–88.

Tong, Y. L. (1990) *The multivariate normal distribution.* Springer Series in Statistics. New York etc.: Springer-Verlag.
Wang, J., Gui, L., Su, W., Sabatti, C. and Owen, A. (2021) Detecting multiple replicating signals using adaptive filtering procedures. *Annals of Statistics, accepted for publication*.

Wang, J. and Owen, A. (2019) Admissibility in partial conjunction testing. *Journal of the American Statistical Association, 114*, 158–168.

Westfall, P. and Young, S. (1993) *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley Series in Probability and Mathematical Statistics. New York.

Whitt, W. (1980) Uniform conditional stochastic order. *J. Appl. Probab., 17*, 112–123. URL: [https://doi.org/10.2307/3212929](https://doi.org/10.2307/3212929).

Whitt, W. (1982) Multivariate monotone likelihood ratio and uniform conditional stochastic order. *Journal of Applied Probability, 19*, 695–701.

Zhao, Q., Small, D. S. and Su, W. (2019) Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association, 114*, 1291–1304.
Web-based supporting materials for:
Multiple testing of partial conjunction null hypotheses with conditional $p$-values based on combination test statistics

Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bibliothekstr. 5, D-28359 Bremen, Germany. E-mail:dickhaus@uni-bremen.de

Ruth Heller, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv-Yafo, Israel. E-mail:ruheller@tauex.tau.ac.il

Anh-Tuan Hoang, Institute for Statistics, University of Bremen, Bibliothekstr. 5, D-28359 Bremen, Germany. E-mail:anhtuan.hoang@uni-bremen.de

S1 The empirical distribution of the null PC $p$-values before and after conditioning on selection

When testing a family of PC hypotheses, it is common that the majority of the PC null hypotheses $p$-values may be very conservative (i.e., they have a distribution that is stochastically larger than uniform). This follows since in practice many of the true $H^{\gamma/s}$ hypotheses have signal in less than $\gamma - 1$ studies. Even if the $p$-value of each true individual null hypothesis is uniformly distributed, the $p$-value for a true PC hypothesis $H^{\gamma/s}$ can be uniform only if there are $\gamma - 1$ nonnull individual hypotheses and their $p$-values are zero (i.e., there is overwhelming evidence that $\gamma - 1$ studies have effect).

In the setting considered in Section 5, the top row of Figure S1 shows that the distribution of PC null $p$-values is conservative for $\gamma \geq 2$, and the conservativeness is more severe for larger values of $\gamma$: the cummulative distribution function (CDF) is further below the 45-degree diagonal line as $\gamma$ increases. The bottom row of Figure S1 shows that the distribution of the conditional PC null $p$-values is far less conservative, but still conservative.
Figure S1: In the data generation described in Section 4 (symmetric block dependent setting, with block correlation 0.9 in blocks of size 10, with 0.9 true global null hypotheses and 1-0.9-0.002= 0.098 additional true PC null hypotheses distributed evenly in all possible configurations), the estimated cumulative distribution function of the PC null \( p \)-values (row 1) and conditional PC null \( p \)-values following selection with a pre-specified threshold \( \tau = 0.1 \). Columns 1 to 4 correspond to \( \gamma \) values 2 to 5, respectively.

\section*{S2 Reduced dependence across PC \( p \)-values}

Since each PC \( p \)-value is formed by combining \( s \) independent studies, we expect the dependence across PC \( p \)-values to be smaller than within individual studies with dependent test statistics. Moreover, since only the top \( s - \gamma + 1 \) \( p \)-values are combined, we expect the dependence to be weaker as \( \gamma \) increases. This is confirmed for the symmetric block correlation setting in Section 5 in Figure S2. To generate this figure, we restricted ourselves to combining standard normal \( z \)-scores within each study into PC \( p \)-values, and then transformed the PC \( p \)-values into \( z \)-scores using the Gaussian quantile function. For two features that have correlation value \( \rho \in \{0.1, \ldots, 0.9\} \) within each study (i.e., they are within the same block), we computed the correlation of the PC \( z \)-scores. We see that the PC \( z \)-scores are less correlated than the \( Z \)-scores within each study, and that the correlation decreases as \( \gamma \) increases.
Figure S2: In the data generation described in Section 4 (symmetric block dependent setting, with blocks of size 10) the correlation between two PC Z-scores from the same block versus the correlation between two PC Z-scores from the same block in the same study for $\gamma = 3$ (solid black), $\gamma = 5$ (dashed black), and $\gamma = 7$ (dotted black). The 45 degrees diagonal line is in red.
Asymptotic FDR control when $m \to \infty$

In high dimensional applications, $p$-values are typically not independent. Storey et al. (2004) assumed the following for asymptotic FDR control on a family of $m$ hypotheses:

$$\forall t \in (0, 1] : \lim_{m \to \infty} \frac{V_m(t)}{m_0} = G_0(t) \text{ and } \lim_{m \to \infty} \frac{R_m(t) - V_m(t)}{m - m_0} = G_1(t) \text{ a.s.}, \tag{S3.1}$$

where $m_0$ is the number of true null hypotheses; $V_m(t)$ is the (random) number of true null hypotheses with $p$-values below $t$; $R_m(t)$ is the number of (random) $p$-values below $t$; and $G_0$ and $G_1$ are continuous functions such that

$$\forall t \in (0, 1] : 0 < G_0(t) \leq t; \tag{S3.2}$$

$$\lim_{m \to \infty} \frac{m_0}{m} = \pi_0 \text{ exists.} \tag{S3.3}$$

In their Theorem 6, they prove that if the convergence assumptions in (S3.1)–(S3.3) hold, then for each $\delta > 0$,

$$\lim_{m \to \infty} \inf_{t \geq \delta} \left\{ \frac{\hat{\pi}_0(\lambda)t}{\{R_m(t) \vee 1\}/m} - \frac{V_m(t)}{R_m(t) \vee 1} \right\} \geq 0$$

holds almost surely, where $\hat{\pi}_0(\lambda) = \frac{m - R_m(\lambda)}{(1 - \lambda)m}$ is the plug-in estimate for the number of true null hypotheses. The implication of this result is that asymptotically, the false discovery proportion (FDP) is almost surely at most $\alpha$ if the adaptive BH procedure at level $\alpha$ is applied on the $m$ $p$-values (so the FDR, which is the expected FDP, is also controlled at level $\alpha$ asymptotically.)

In many high dimensional applications, the dependence within the study is indeed local, thus it is reasonable to assume the convergence assumptions. This is the case in chief genomics applications: for GWAS or eQTLs, the covariance structure is that of a banded matrix; for microarrays or RNA-seq experiments, gene-gene networks are typically sparse (Wang et al., 2021).

We provide similar guarantees of our methodology, when inferring on multiple studies with local
dependence. For the family of PC hypotheses, using the adaptive BH procedure on the PC p-
values provides asymptotic FDP (and FDR) control if the convergence assumptions (S3.1)–(S3.3)
are satisfied for the data generating the PC p-values. This is guaranteed, for example, if within
each study there is a single null distribution from which the p-values from true null hypotheses
are generated, and a single nonnull distribution form which the p-values from false null hypotheses
are generated, and the limit for the fraction of each of the 2s combinations of null and nonnull
hypotheses exists. More specifically, let \( \mathbf{h} \) be the vector of hypotheses status indicators for a feature
(so \( h_i = 1 \) if the \( i \)th null hypothesis is true, and zero otherwise, for \( i = 1, \ldots, s \)). Then if for every \( \mathbf{h} \),
the limit for the fraction of features with hypothesis states \( \mathbf{h} \) exists, then convergence assumptions
(S3.1)–(S3.3) are satisfied.

Our algorithm, in which we first select the PC hypotheses which have PC p-values at most the
selection threshold \( \tau \), and then applies the BH procedure on the conditional PC p-values, also
provides asymptotic FDP (and FDR) control. This result follows from the following proposition.

**Proposition S3.1.** Assume that condition (A1) or (A2) is satisfied for each p-value. More-
over, assume that the convergence assumptions of equations (S3.1)–(S3.3) hold for the PC p-values
\( \{p_j^{\gamma/s}, j = 1, \ldots, m\} \). Then, the FDP of the BH procedure at level \( \alpha \) on \( \{p_j^{\gamma/s}/\tau, j \in S_\tau\} \), for a fixed
pre-specified \( \tau > 0 \), is asymptotically almost surely at most \( \alpha \).

**Proof.** The (random) BH threshold, computed on the set of PC p-values which are considered in
step (iv) of our Algorithm 2.1, is

\[
\hat{x} = \max \left\{ x : \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} \leq \alpha \right\},
\]

(S3.4)

where \( |S_{\tau x}| = \sum_{j \in S_\tau} \mathbb{I} \left( \frac{p_j^{\gamma/s}}{\tau} \leq x \right) = \sum_{j=1}^m \mathbb{I} \left( p_j^{\gamma/s} \leq \tau x \right) \) is the number of rejected hypotheses when
thresholding the conditional PC p-values at level \( x \). So it is enough to show that

\[
\lim_{m \to \infty} \sup_{0 \leq x \leq 1} \left\{ FDP(\tau x) - \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} \right\} \leq 0
\]
almost surely, where \( FDP(\tau x) \) is the false discovery proportion when the hypotheses with PC \( p \)-values at most \( \tau x \) are rejected.

Let \( \mathcal{H}^{\gamma/s} \) be the set of indices of the true PC null hypotheses (so \( |\mathcal{H}^{\gamma/s}| = m_0 \)). From Theorem 1, it follows that if \( \mathcal{H}_0^{\gamma/s} \) is true, then \( \mathbb{P}(P_{\gamma/s}^j / \tau \leq x) \leq x \mathbb{P}(P_{\gamma/s}^j \leq \tau) \) for all \( x \in [0,1] \). Therefore,

\[
G_0(\tau x) = \lim_{m \to \infty} \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} I(p_{\gamma/s}^j \leq \tau x)}{|\mathcal{H}^{\gamma/s}|} = \mathbb{E} \left( \lim_{m \to \infty} \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} I(p_{\gamma/s}^j \leq \tau x)}{|\mathcal{H}^{\gamma/s}|} \right) = \lim_{m \to \infty} \mathbb{E} \left( \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} I(p_{\gamma/s}^j \leq \tau x)}{|\mathcal{H}^{\gamma/s}|} \right) = \lim_{m \to \infty} \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} \mathbb{P}(p_{\gamma/s}^j \leq \tau x)}{|\mathcal{H}^{\gamma/s}|} \leq \lim_{m \to \infty} \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} \mathbb{P}(p_{\gamma/s}^j \leq \tau x)}{|\mathcal{H}^{\gamma/s}|} = G_0(\tau x). \tag{S3.5}
\]

From (a slight modification) of the Glivenko-Cantelli theorem (Storey et al., 2004),

\[
\lim_{m \to \infty} \sup_{0 \leq t \leq 1} \left| \frac{|S_t|}{m} - \pi_0 G_0(t) - (1 - \pi_0) G_1(t) \right| = 0
\]

almost surely. Therefore,

\[
\lim_{m \to \infty} \sup_{0 \leq x \leq 1} \left| \frac{|S_{\tau x}| \times x}{|S_{\tau x}| \vee 1} - \frac{\pi_0 G_0(\tau x) x + (1 - \pi_0) G_1(\tau x) x}{(\pi_0 G_0(\tau) x + (1 - \pi_0) G_1(\tau x)) \vee 1} \right| = 0 \tag{S3.6}
\]

almost surely.

We use the above results in order to show that the difference between the FDP and its BH point estimate is at most zero. We start by expressing this difference as three differences, and then argue
that each of these differences is a.s. upper bounded by zero:

\[ FDP(\tau x) - \frac{|S_\tau|x}{|S_\tau|} \lor 1 = \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \leq \tau x) \lor 1}{\sum_{j=1}^{m} \mathbb{I}(p_j^{\gamma/s} \leq \tau x)} - \frac{|S_\tau|x}{|S_\tau|} \lor 1 = \sum_{j \in \mathcal{H}} \gamma/s I(p_j^{\gamma/s} \leq \tau x) \lor 1 \sum_{m} \mathbb{I}(p_j^{\gamma/s} \leq \tau x) \]

\[ = \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)} - \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)} \]

\[ + \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)} - \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)} \]

\[ (S3.7) \]

Result follows since: from assumptions (S3.1)–(S3.3), as in Storey et al. (2004), almost surely

\[ \lim_{m \to \infty} \sup_{0 \leq x \leq 1} \left| \frac{\sum_{j \in \mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \leq \tau x) \lor 1}{(\sum_{j=1}^{m} \mathbb{I}(p_j^{\gamma/s} \leq \tau x)) \lor 1} - \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)) \lor 1} \right| = 0; \]

from (S3.5),

\[ \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)) \lor 1} - \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)) \lor 1} \leq 0; \]

and from (S3.6),

\[ \lim_{m \to \infty} \sup_{0 \leq x \leq 1} \left\{ \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1 - \pi_0) G_1(\tau x)) \lor 1} - \frac{|S_\tau|x}{|S_\tau| \lor 1} \right\} \leq 0. \]

Our next goal is to generalize Proposition S3.1 to cases in which the value of \( \tau \) is selected on the basis of the available data. Thus, in Proposition S3.2 below we consider a random variable \( \hat{\tau} \) which describes the selection rule, meaning that \( \tau = \hat{\tau}(\text{data}) \).

**Proposition S3.2.** Assume that condition (A1) or (A2) is satisfied for each p-value. More-
over, assume that the convergence assumptions of equations (S3.1)–(S3.3) hold for the PC p-values \(\{p_{\gamma/s}^j, j = 1, \ldots, m\}\).

For any given value \(\tau \in (0, 1]\), define the (random) function \(\hat{F}_\tau : [0, 1] \to [0, 1]\) by \(\hat{F}_\tau(x) = |S_{\tau x}|/|S_\tau|\).

Let \(\hat{\tau} \equiv \hat{\tau}_m\) denote a (0,1]-valued random variable which is measurable with respect to (the \(\sigma\)-field generated by) the available (random) data. Assume that the sequence \(\{\hat{\tau}_m\}_{m \geq 1}\) possesses an almost sure limiting value \(\tau_\infty \in (0, 1]\), and that \(||\hat{F}_{\tau_m} - \hat{F}_{\tau_\infty}||_\infty \to 0\) almost surely as \(m \to \infty\).

Then, the FDP of the BH procedure at level \(\alpha\) on \(\{p_{\gamma/s}^{\hat{\tau}_m}/\hat{\tau}_m, j \in S_{\hat{\tau}_m}\}\) is asymptotically almost surely at most \(\alpha\).

Proof. Consider the following representation of the BH threshold \(\hat{x} \equiv \hat{x}(\tau)\) for a given \(\tau\) (and a given \(\alpha\)):

\[
\hat{x}(\tau) = \max \left\{ x \in (0, 1] : \frac{x}{\hat{F}_\tau(x)} \geq \frac{\alpha}{x} \right\}, \tag{S3.8}
\]

if the maximum in (S3.8) exists, and \(\hat{x}(\tau) = 0\) otherwise (see, e.g., Lemma 5.7 in Dickhaus (2014)). The graph of the function \(x \mapsto x/\alpha\) appearing in (S3.8) is occasionally referred to as the "Simes line". Proposition S3.1 then yields (under the stated assumptions), that choosing \(\hat{x}(\tau)\) as the rejection threshold for the conditional PC p-values leads to an FDP which is asymptotically almost surely upper-bounded by \(\alpha\) for any fixed \(\tau\). In particular, considering \(\tau = \tau_\infty\) in (S3.8) leads to an FDP which is asymptotically almost surely upper-bounded by \(\alpha\), because \(\tau_\infty\) is a fixed constant in the interval \((0, 1]\), by assumption.

Analogously, the BH threshold for a random (selected) value \(\hat{\tau}_m\) is given by

\[
\hat{x}(\hat{\tau}_m) = \max \left\{ x \in (0, 1] : \frac{x}{\hat{F}_{\hat{\tau}_m}(x)} \leq \alpha \right\}, \tag{S3.9}
\]
if the maximum in (S3.9) exists, and \( \hat{x}(\hat{\tau}_m) = 0 \) otherwise. Clearly, this representation implies that \( \hat{x}(\hat{\tau}_m) \) depends on the data only via \( \hat{F}_{\hat{\tau}_m} \), as soon as \( \hat{\tau}_m \) has been chosen. By our assumptions, \( \hat{\tau}_m \) converges almost surely to \( \tau_\infty \) and \( |\hat{F}_{\hat{\tau}_m} - \hat{F}_{\tau_\infty}| \) converges uniformly and almost surely to zero. Furthermore, the mapping \( \hat{F}_\tau \mapsto \hat{x}(\tau) \) is continuous. From these assertions, we conclude that \( |\hat{x}(\hat{\tau}_m) - \hat{x}(\tau_\infty)| \) converges to zero almost surely as \( m \) tends to infinity. However, as argued before, choosing \( \hat{x}(\tau_\infty) \) as the rejection threshold for the conditional PC \( p \)-values leads to an FDP which is asymptotically almost surely upper-bounded by \( \alpha \), which yields the assertion of the proposition. \( \square \)

The proof of Proposition 3.3 follows from Proposition S3.2 since \( \hat{\tau}_m \) depends on the data only via \( \{|S_{\tau_i}|/m, \hat{F}_{\tau_i}, i = 1, \ldots, K\} \). Since \( |S_{\tau_i}|/m \) and \( \hat{F}_{\tau_i} \) converge almost surely to well defined limiting functions for \( i = 1, \ldots, K \), there exists a limiting value \( \tau_\infty \in \{\tau_1, \ldots, \tau_K\} \) that \( \hat{\tau}_m \) converges to almost surely, and \( ||\hat{F}_{\hat{\tau}_m} - \hat{F}_{\tau_\infty}||_\infty \to 0 \) almost surely as \( m \to \infty \).

### S4 Additional simulations

#### S4.1 Results using Stouffer and Simes combination \( p \)-values

In the setting considered in Section 5, with \( m = 20000 \) and \( \pi_1 = 0.002 \), we applied the novel procedures using Stouffer and Simes combination \( p \)-values. Tables S1 and S2 show the average number of true discoveries (our measure of power) and FDP for the novel procedures, using the selection threshold pre-specified as \( \tau = 0.1 \) or adaptively chosen as in Zhao et al. (2019) with \( \beta = 0.1, 0.5 \). Compared with the unconditional approach of applying BH on the PC \( p \)-values, we see that the power is greater with the novel approach, for every \( \gamma \). The power advantage over the unconditional approach is very large, especially when the \( p \)-values are combined using Simes method. Combining \( p \)-values using Stouffer is more powerful than using Simes. However, a comparison with Table 1 in the main manuscript shows that the power is highest using the Fisher combining method. Although Fisher combining has better power properties than Stouffer and Simes
Table S1: In the symmetric block dependent setting, the average number of true discoveries for \( \gamma = 2, 3, 4, 5 \) using the Stouffer combining method (rows 1–4) and the Simes combining method (rows 5–8), for the following procedures at level 0.05: BH on PC \( p \)-values, BH and adaptive BH (denoted aBH) on conditional PC \( p \)-values using selection threshold \( \tau = 0.1 \), adaptive threshold at \( \beta = 0.1 \) and adaptive threshold at \( \beta = 0.5 \). Based on 5000 repetitions.

| Combining method | \( \gamma \) | \( \tau = 0.1 \) | \( \hat{\tau} \) with \( \beta = 0.1 \) | \( \hat{\tau} \) with \( \beta = 0.5 \) |
|------------------|-----------|----------------|----------------|----------------|
| Stouffer         | 2         | 17.7           | 20.0           | 19.9           | 19.7           | 20.2           | 20.1           |
|                  | 3         | 9.0            | 12.7           | 12.4           | 12.6           | 12.4           | 12.8           | 12.9           |
|                  | 4         | 4.0            | 7.4            | 7.4            | 7.1            | 7.0            | 7.4            | 7.5            |
|                  | 5         | 1.6            | 4.1            | 4.2            | 3.8            | 3.7            | 4.0            | 4.0            |
| Simes            | 2         | 6.3            | 17.9           | 17.3           | 18.1           | 18.3           | 18.5           | 19.2           |
|                  | 3         | 0.5            | 7.2            | 7.8            | 6.7            | 6.5            | 6.9            | 7.6            |
|                  | 4         | 0.0            | 1.5            | 1.7            | 1.3            | 1.1            | 1.4            | 1.4            |
|                  | 5         | 0.0            | 0.3            | 0.4            | 0.3            | 0.2            | 0.3            | 0.3            |

in many data generation settings, this is not always true: Wang and Owen (2019) showed that any combining function increasing in its coordinates, that uses the \( s - \gamma + 1 \) largest \( p \)-values to provide a valid PC \( p \)-value, is admissible.

From Table S2 it is clear that the conditional and unconditional procedures are below the nominal 0.05 level, that the unconditional approach is the most conservative (i.e., with lowest FDR level), and that using Simes (rather than Stouffer or Fisher) for combining is most conservative.

S4.2 Results with strong dependence among the \( p \)-values

In the main manuscript we showed results for weak dependence using a dependence block size of 10, and 2000 independent blocks, for a total of \( m = 20000 \) PC null hypotheses. In this section we alter the block size to 100, keeping all other configurations of the data generation the same. This is a setting with strong dependence, since the blocks are large and we only have 200 independent blocks, for a total of \( m = 20000 \) PC null hypotheses.

Tables S3 and S4 show the average number of true discoveries and FDP for the novel procedures, us-
Table S2: In the symmetric block dependent setting, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$, using the Stouffer combining method (rows 1–4) and the Simes combining method (rows 5–8), for the following procedures at level 0.05: BH on PC $p$-values, BH and adaptive BH (denoted aBH) on conditional PC $p$-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

| Combining method | $\gamma$ | BH | BH | aBH | BH | aBH | BH | aBH |
|-----------------|---------|----|----|-----|----|-----|----|-----|
|                 | 2       | 0.005 | 0.016 | 0.014 | 0.015 | 0.014 | 0.017 | 0.017 |
|                 | 3       | 0.003 | 0.017 | 0.016 | 0.017 | 0.016 | 0.019 | 0.020 |
|                 | 4       | 0.001 | 0.020 | 0.020 | 0.017 | 0.016 | 0.019 | 0.021 |
|                 | 5       | 0.000 | 0.016 | 0.017 | 0.013 | 0.012 | 0.015 | 0.016 |
| Simes           | 2       | 0.001 | 0.011 | 0.010 | 0.012 | 0.012 | 0.012 | 0.014 |
|                 | 3       | 0.000 | 0.006 | 0.007 | 0.005 | 0.005 | 0.006 | 0.007 |
|                 | 4       | 0.000 | 0.001 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 |
|                 | 5       | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

ing the selection threshold pre-specified as $\tau = 0.1$ or adaptively chosen as in Zhao et al. (2019) with $\beta = 0.1, 0.5$, versus competitors. The average number of true discoveries and FDP are remarkably similar to the averages in Tables 1 and 2 in the main manuscript for adaFilter, the unconditional approach, and the conditional approach with $\tau = 0.1$. However, the power of the novel procedures with the selection threshold adaptively chosen have lower power compared with the weak dependence setting in the main manuscript. Table S5 shows the reason for the power deterioration: when the dependence is stronger, the estimated selection threshold tends to be higher than for weak dependence.

S4.3 Results using the adaptive method of Hoang and Dickhaus (2021b,a) for estimating $\pi_0$

Storey’s estimator (Storey et al., 2004) for the fraction of null hypotheses tends to be conservative if the $p$-values from null hypotheses have a distribution that is stochastically larger than uniform. Therefore, Hoang and Dickhaus (2021b,a) suggested to randomize the $p$-values first, and then apply Storey’s estimation method on the vector of randomized (rather than original) $p$-values.
Table S3: In the symmetric block dependent setting with block size 100, the average number of true discoveries for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC $p$-values, BH and adaptive BH (denoted aBH) on conditional PC $p$-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

| $\pi_1$ | $\gamma$ | adaFilter | BH | conditional | BH | aBH | BH | aBH | BH | aBH |
|---------|----------|-----------|----|-------------|----|-----|----|-----|----|-----|
|         |          |           |    | $\tau=0.1$ |    |     |    |     |    |     |
| 0.002   | 2        | 25.4      | 25.8| 28.6        | 28.4| 26.8 | 26.3| 27.5 | 27.1 |
|         | 3        | 16.9      | 14.4| 20.0        | 20.1| 17.7 | 16.8| 18.6 | 18.0 |
|         | 4        | 10.5      | 6.2 | 12.0        | 12.3| 10.0 | 9.2 | 10.8 | 10.2 |
|         | 5        | 6.9       | 2.0 | 6.3         | 6.5 | 4.9  | 4.3 | 5.4  | 4.9  |
| 0.02    | 2        | 338.7     | 302.4| 320.1       | 328.4| 312.3 | 307.4| 316.7 | 314.3 |
|         | 3        | 281.5     | 202.4| 242.8       | 269.7| 238.7 | 231.0| 245.5 | 244.1 |
|         | 4        | 218.1     | 110.1| 160.3       | 196.1| 165.7 | 158.7| 172.1 | 172.3 |
|         | 5        | 168.7     | 46.7 | 91.8        | 123.7| 104.6 | 100.0| 109.7 | 112.5 |

Table S4: In the symmetric block dependent setting with block size 100, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC $p$-values, BH and adaptive BH (denoted aBH) on conditional PC $p$-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

| $\pi_1$ | $\gamma$ | adaFilter | BH | conditional | BH | aBH | BH | aBH | BH | aBH |
|---------|----------|-----------|----|-------------|----|-----|----|-----|----|-----|
|         |          |           |    | $\tau=0.1$ |    |     |    |     |    |     |
| 0.002   | 2        | 0.046     | 0.003| 0.012       | 0.013| 0.006 | 0.004| 0.008 | 0.007 |
|         | 3        | 0.044     | 0.001| 0.014       | 0.015| 0.005 | 0.004| 0.009 | 0.007 |
|         | 4        | 0.043     | 0.001| 0.013       | 0.015| 0.005 | 0.004| 0.008 | 0.006 |
|         | 5        | 0.043     | 0.000| 0.010       | 0.011| 0.004 | 0.003| 0.006 | 0.005 |
| 0.02    | 2        | 0.034     | 0.004| 0.010       | 0.017| 0.007 | 0.005| 0.009 | 0.008 |
|         | 3        | 0.031     | 0.002| 0.007       | 0.019| 0.007 | 0.005| 0.008 | 0.008 |
|         | 4        | 0.027     | 0.001| 0.006       | 0.016| 0.007 | 0.006| 0.008 | 0.008 |
|         | 5        | 0.022     | 0.000| 0.003       | 0.009| 0.005 | 0.004| 0.006 | 0.007 |
Table S5: In the symmetric block dependent setting with block size 100, the average adaptively selected $\tau$ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

| $\pi_1$ | $\gamma$ | $\tilde{\tau}$ with $\beta = 0.1$ | $\tilde{\tau}$ with $\beta = 0.5$ |
|---------|---------|-------------------------------|-------------------------------|
| 0.002   | 2       | 0.63                          | 0.40                          |
|         | 3       | 0.52                          | 0.35                          |
|         | 4       | 0.54                          | 0.41                          |
|         | 5       | 0.57                          | 0.46                          |
| 0.02    | 2       | 0.64                          | 0.43                          |
|         | 3       | 0.55                          | 0.40                          |
|         | 4       | 0.56                          | 0.45                          |
|         | 5       | 0.59                          | 0.50                          |

The method of Hoang and Dickhaus (2021b,a) can provide a better estimate of the fraction of PC null hypotheses among the selected than Storey’s estimator (Storey et al., 2004), in settings where the conditional PC $p$-values are still stochastically larger than uniform. We set out to examine this in our simulation settings. Both methods were implemented using the fixed parameter $\lambda = 0.5$. Specifically, Storey’s estimate is $2 \times \left( \sum_{i \in S,} \mathbb{I} \left( \frac{p_i^{\gamma/s}}{\tau} > 0.5 \right) + 1 \right)$; the estimator of Hoang and Dickhaus (2021b,a) is similar, except $\frac{p_i^{\gamma/s}}{\tau}$ is replaced by

$$U \mathbb{I} \left\{ \frac{p_i^{\gamma/s}}{\tau} > 0.5 \right\} + 2 \times \frac{p_i^{\gamma/s}}{\tau} \mathbb{I} \left\{ \frac{p_i^{\gamma/s}}{\tau} \leq 0.5 \right\},$$

where $U$ is a Uni[0, 1] random variable.

Figure S3 shows that both estimation methods are conservative, and tend to overestimate the fraction of PC null hypotheses among the selected. However, for $\gamma = 2, 3$ the method of Hoang and Dickhaus (2021b,a) is far less conservative. For $\gamma = 4, 5$, there is little difference between the methods. Table S6 shows the gain in power from using each method.
Figure S3: The estimated fraction of PC null hypotheses, using the plug-in method of Storey et al. (2004) (red triangles) and using the method of Hoang and Dickhaus (2021b,a) (black circles), versus the true fraction of PC null hypotheses among the selected. Plotted are the results from 5000 data generations from the symmetric block dependence described in Section 5, with $\pi_1 = 0.02$. The threshold for selection was $\tau = 0.25$. The black line is the 45 degree diagonal line. Each panel is a different $\gamma$. 
Table S6: In the symmetric block dependent setting with $\pi_1 = 0.02$, for $\gamma = 2, 3, 4, 5$, the average number of true rejections using our conditional approach with selection threshold $\tau = 0.25$, when the multiple testing procedure on the conditional PC $p$-values is: BH (column 4); adaptive BH using Storey’s plug-in estimate (column 5); and adaptive BH using the method of Hoang and Dickhaus (2021b,a) (column 6). Columns 2 and 3 provide the average value for the estimated fraction of PC nulls among the selected, using Storey’s plug-in estimate and the method of Hoang and Dickhaus (2021b,a), respectively. Based on 5000 simulations.

| $\gamma$ | Storey’s $\hat{\pi}_0$ | Hoang and Dickhaus’s $\hat{\pi}_0$ | BH       | adaptive BH Storey’s $\hat{\pi}_0$ | adaptive BH Hoang and Dickhaus’s $\hat{\pi}_0$ |
|---------|----------------|----------------|---------|---------------------------------|---------------------------------|
| 2       | 1.02           | 0.88           | 320.30  | 319.88                         | 322.75                         |
| 3       | 0.85           | 0.73           | 250.38  | 255.69                         | 260.89                         |
| 4       | 0.68           | 0.65           | 176.03  | 192.79                         | 195.01                         |
| 5       | 0.60           | 0.64           | 109.62  | 133.62                         | 130.54                         |

S5  A different approach of choosing $\tau$ adaptively

As mentioned in Section 3.1, one can consider any function $G(\tau) = H(|S_\tau|, \tau)$, where $H$ is increasing in $|S_\tau|$ and decreasing in $\tau$, and minimize $G$. The idea is always to minimize the number $|S_\tau|$ of non-discarded $p$-values while trying to keep the penalty $1/\tau$ small. For example, minimize $G(\tau) = |S_\tau| - km\tau$, where $m$ is the number of $p$-values, and $k \in (0, \infty)$ is a pre-specified constant.

Since $|S_\tau| = m\hat{F}_m(\tau)$, we approximate $G'(\tau)$ with

$$m \left[ \frac{\hat{F}_m(\tau + \omega) - \hat{F}_m(\tau)}{\omega} - k \right],$$

where $\hat{F}_m$ is the ecdf of the $p$-values, and $\omega > 0$ is a small pre-specified constant.

Given a sequence of $\tau$’s, $\tau_1 < \ldots < \tau_K$, we go from $\tau_i$ to $\tau_{i-1}$, starting with $\tau_K$, if there is sufficient evidence that $G'(\tau_i) > 0$. We can compare $\hat{F}_m(\tau_i + \omega) - \hat{F}_m(\tau_i)$ and $k\omega$ directly or we can try to reject $q > k$, where $m(\hat{F}_m(\tau_i + \omega) - \hat{F}_m(\tau_i)) \sim \text{Binomial}(m, q\omega)$ at a pre-specified significance level $\beta$ (cf. Zhao et al. (2019)).

Graphically, if $k = 1$, minimizing $G$ is maximizing the vertical distance between the identity line and the ecdf $\hat{F}_m$ (below the identity line), which is equivalent to maximizing the distance between
Figure S4: Maximizing the distance (below the identity line) between the identity line and the ecdf $\hat{F}_m$ of the $p$-values. The length of the vertical line is $-G$.

the identity line and the ecdf $\hat{F}_m$, see Figure S4.

S6 Selection before combination

As described in Algorithm 2.1, we select in step (ii) from the $p$-values $P_j^{\gamma/s}$, $j = 1, \ldots, m$, i.e. after combining the base $p$-values $P_{1,j}, \ldots, P_{s,j}$. According to Theorem 1, if the latter fulfill either (A1) or (A2), the partial conjunction $p$-value $P_j^{\gamma/s}$ is uniformly valid. However, both (A1) and (A2) mean that $P_{1,j}, \ldots, P_{s,j}$ are uniformly valid themselves. Thus, the conditional $p$-values $(P_{i,j}/\tau_{i,j}, | P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$, are valid for $H_{j,1}, \ldots, H_{j,s}$, respectively, for any $\tau_{i,j} \in [0,1]$, $i = 1, \ldots, s$. For the sake of simplicity we set $\tau_{i,j} = \tau^{(1)}$ for all $i, j$.

Furthermore, it is easy to see that $(P_{i,j}/\tau_{i,j}, | P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$ satisfy (A1) or (A2) if $P_{1,j}, \ldots, P_{s,j}$ do, respectively. Therefore, one can condition again, after combining $(P_{i,j}/\tau_{i,j}, | P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$ for the
partial conjunction null hypothesis $H_j^{\gamma/S_j,\tau^{(1)}}$, where $S_{j,\tau^{(1)}} = \#\{P_{i,j}/\tau^{(1)}, i = 1, \ldots, s \mid P_{i,j} \leq \tau^{(1)}\}$.

In case of $\gamma > S_{j,\tau^{(1)}}$, we retain $H_j^{\gamma/s}$.

There are two sources of conservativity when dealing with partial conjunction p-values. The base, null p-values $P_{1,j}, \ldots, P_{s,j}$ can be stochastically larger than Uni[0,1], or the partial conjunction p-value $P_j^{\gamma/s}$ can be conservative due to the nature of partial conjunction null hypotheses. In the paper we deal with the second kind directly and with the first kind only indirectly.

For this section, we call $\tau_1 = \cdots = \tau_m = \tau^{(2)}$ the selection parameter applied after combining (this was called $\tau$ in the paper). For fixed $\tau^{(1)}, \tau^{(2)}$, let $P_{j,0}^{0,0} = P_j^{\gamma/s}$, $P_{j,1}^{1,0}$, $P_{j,1}^{0,1}$, $P_{j,1}^{1,1}$, be the unconditional p-values, the selected p-values after conditioning with $\tau^{(1)}$ only before combining, the selected p-values after conditioning with $\tau^{(2)}$ only after combining, and the selected p-values after conditioning with $\tau^{(1)}$ and $\tau^{(2)}$ before and after combining, respectively. The index sets $S_{0,0}, S_{1,0}, S_{0,1}, S_{1,1}$ comprise of the selected indices $j \in \{1, \ldots, m\}$. In the paper, we only consider $P_{j,0}^{0,0}$ and $P_{j,1}^{0,1}$, where $S_{0,0} = \{1, \ldots, m\}$ and $S_{0,1} = S_{\tau^{(2)}}$.

For our simulations, we set $\tau^{(1)} = 0.7$ and $\tau^{(2)} = 0.5$. For our analysis, we consider Model 1 from Section 4, where $\theta_{i,j}^* = 0$ and $n_{i,j} = 50$ for all $i, j$. We consider $s = 6, \gamma = 3$ for a fixed endpoint $j$, parameter values $\theta_j = (\theta_{1,j}, \ldots, \theta_{s,j}) \in H_j^{\gamma/s}$, and $P_j^{\gamma/s}$ the Fisher combination. Via Monte-Carlo simulations with 100,000 repetitions, we approximate the probability for $j$ to be selected, $j \in S_{0,0}, S_{1,0}, S_{0,1}, S_{1,1}$, and, if selected, approximate the cdfs of $P_{j,0}^{0,0}, P_{j,1}^{1,0}, P_{j,1}^{0,1}, P_{j,1}^{1,1}$. We consider the parameter values $\theta_j$ as described in Table S7. Under the (LFC-) parameter value $\theta_j = (0, 0, 0, 0, -\infty, -\infty)$, $P_{j}^{3/6}$ is uniformly distributed, so $P_{j}^{3/6}$ is close to Uni[0,1] under Setting (0,0). Under Setting (1,0), the conservativity of $P_{j}^{3/6}$ comes from the conservative base p-values $P_{1,j}$ and $P_{2,j}$, whereas under Setting (0,1), $P_{j}^{3/6}$ is conservative due to the number of base, null p-values.

In the last setting both sources of conservativity are present. The results are in Figure S5.

Between $P_{j,1}^{1,0}$ and $P_{j,1}^{0,1}$, the first works better in Setting (1,0) and the second in Setting (0,1). The unconditional p-value $P_{j,0}^{0,0}$ produces the most conservative p-values in all settings, and the
Figure S5: Approximations of the cdfs of the (null) $p$-values $P^{0,0}$, $P^{1,0}$, $P^{0,1}$, $P^{1,1}$, given that the $p$-value is selected under different settings with different sources of conservativity.
Table S7: The considered parameter values $\theta_j$ in the simulations.

| Setting | $\theta_{1,j}\sqrt{n_{1,j}}$ | $\theta_{2,j}\sqrt{n_{1,j}}$ | $\theta_{3,j}\sqrt{n_{1,j}}$ | $\theta_{4,j}\sqrt{n_{1,j}}$ | $\theta_{5,j}\sqrt{n_{1,j}}$ | $\theta_{6,j}\sqrt{n_{1,j}}$ |
|---------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| (0,0)   | 0                             | 0                             | 0                             | 0                             | -5                            | -5                            |
| (1,0)   | 1                             | 1                             | 0                             | 0                             | -5                            | -5                            |
| (0,1)   | 0                             | 0                             | 0                             | 0                             | 0                             | -5                            |
| (1,1)   | 1                             | 1                             | 1                             | 0                             | 0                             | -5                            |

$p$-value $P_{j}^{1,1}$ is left with the least conservative $p$-values, after selection, in almost each setting. The probability of selecting $j$ is lowest for $P_{0}^{0,1}$ and $P_{j}^{1,1}$. Note that the discarding event $j \notin S_{1,0}$ happens if and only if $\gamma > S_{j,\tau(1)}$.

S7 Regarding the combination $p$-value $P_{\gamma/s}$

We assumed that the combination function $P_{\gamma/s}(P_{1}, \ldots, P_{s})$ is increasing in each $p$-value $P_{1}, \ldots, P_{s}$ and valid for $H_{\gamma/s}$. Since it holds $H_{1}^{s} \subseteq H_{2}^{s} \subseteq \cdots \subseteq H_{s}^{s}$, valid $p$-values for $H_{\gamma}^{1}$ need not be valid for $H_{\gamma}^{s}$, $\gamma < \gamma'$. However, valid $p$-values for $H_{\gamma}^{1}$ can be easily derived from (increasing) combination functions for $H_{s-\gamma+1}^{1}$ by essentially combining the $s-\gamma+1$ largest $p$-values, as shown by Benjamini and Heller (2008). For instance, the Fisher $p$-value

$$P_{1/(s-\gamma+1)}^{1}(P_{1}, \ldots, P_{s-\gamma+1}) = 1 - \Phi \left( \frac{1}{\sqrt{s-\gamma+1}} \sum_{i=1}^{s-\gamma+1} \Phi^{-1} (1 - P_{i}) \right)$$

which is valid for $H_{1/(s-\gamma+1)}^{1}$ implies that

$$P_{\gamma/s}(P_{1}, \ldots, P_{s}) = 1 - \Phi \left( \frac{1}{\sqrt{s-\gamma+1}} \sum_{i=\gamma}^{s} \Phi^{-1} (1 - P_{(i)}) \right)$$

is also valid for $H_{\gamma/s}^{s}$.

Furthermore, we assumed that $P_{\gamma/s}$ is uniform under the LFC $\pi(0, \ldots, 0, U_{1}, \ldots, U_{s-\gamma+1})$ in $H_{\gamma/s}$, where $\pi$ is any permutation vector of $s$ elements. This assumption is sufficient (together with (A1) or (A2)) to imply that $P_{\gamma/s}(P_{1}, \ldots, P_{s})$ is a uniformly valid $p$-value for $H_{\gamma/s}^{s}$, as claimed in
Theorem 1. This is, for example, fulfilled if $P_{\gamma/s}$ is derived from a combination $p$-value $P_1/(s-\gamma+1)$, as described above, that is uniform under the LFC $(U_1, \ldots, U_{s-\gamma+1})$. As mentioned in Remark 2, the assumption is not necessary, and

$$\text{Uni}[0,1] \leq_{th} P_{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}))$$

is also sufficient.

We give an example of a valid combination $p$-values $P_{\gamma/s}$ that fulfills neither condition, and one that only fulfills the less strict one. We choose any $s$ and $\gamma$, so that $s - \gamma + 1 = 10$ and consider the harmonic mean and the arithmetic mean, multiplied with constants such, that they are valid for $H_{\gamma/s}$. We approximated their cdfs under an LFC in $H_{\gamma/s}$ with a Monte-Carlo Simulation with 100,000 repetitions, see Figure S6. Graphically, for a given parameter in the null, one can determine if a $p$-value is valid, if its cdf under that parameter is always below the identity line. It is uniformly valid, if the line that connects $(0, 0)$ with the point $(\tau, F(\tau))$ is always above the cdf $F$, for all $\tau$. If a $p$-value is valid under an LFC in the null, then it is valid everywhere in the null, thus the two $p$-values in Figure S6 are indeed valid for $H_{\gamma/s}$. However, we can see that the harmonic mean is not uniformly valid under the LFC whereas the arithmetic mean is. This means that the first, used with base $p$-values that fulfill (A1) or (A2), is not necessarily uniformly valid, but the second is.
Figure S6: The cdfs of the (adjusted) harmonic mean and the (adjusted) arithmetic mean under the LFC $\pi(0, \ldots, 0, U_1, \ldots, U_{10})$ in $H^{\gamma/\gamma+9}$, approximated by a Monte-Carlo simulation with 100,000 repetitions.

References

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

Hoang, A.-T. and Dickhaus, T. (2021a) On the usage of randomized p-values in the Schweder-Spjøtvoll estimator. *Forthcoming in the Annals of the Institute of Statistical Mathematics*.

— (2021b) Randomized p-values for multiple testing and their application in replicability analysis. *Biometrical Journal*, early view.

Storey, J., Taylor, J. and Siegmund, D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66, 187–205.

Wang, J., Gui, L., Su, W., Sabatti, C. and Owen, A. (2021) Detecting multiple replicating signals using adaptive filtering procedures. *Annals of Statistics, accepted for publication*.
Wang, J. and Owen, A. (2019) Admissibility in partial conjunction testing. *Journal of the American Statistical Association*, **114**, 158–168.

Zhao, Q., Small, D. S. and Su, W. (2019) Multiple testing when many \( p \)-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association*, **114**, 1291–1304.