Electronic Journal of Statistics
Vol. 14 (2020) 4244–4272
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1771

Controlling the false discovery exceedance for heterogeneous tests

Sebastian Döhler and Etienne Roquain

e-mail: etienne.roquain@upmc.fr

Abstract: Several classical methods exist for controlling the false discovery exceedance (FDX) for large-scale multiple testing problems, among them the Lehmann-Romano procedure (Lehmann and Romano 2005) ([LR] below) and the Guo-Romano procedure (Guo and Romano 2007) ([GR] below). While these two procedures are the most prominent, they were originally designed for homogeneous test statistics, that is, when the null distribution functions of the p-values F_i, $1 \leq i \leq m$, are all equal. In many applications, however, the data are heterogeneous which leads to heterogeneous null distribution functions. Ignoring this heterogeneity induces a lack of power. In this paper, we develop three new procedures that incorporate the F_i's, while maintaining rigorous FDX control. The heterogeneous version of [LR], denoted [HLR], is based on the arithmetic average of the F_i's, while the heterogeneous version of [GR], denoted [HGR], is based on the geometric average of the F_i's. We also introduce a procedure [PB], that is based on the Poisson-binomial distribution and that uniformly improves [HLR] and [HGR], at the price of a higher computational complexity. Perhaps surprisingly, this shows that, contrary to the known theory of false discovery rate (FDR) control under heterogeneity, the way to incorporate the F_i's can be particularly simple in the case of FDX control, and does not require any further correction term. The performances of the new proposed procedures are illustrated by real and simulated data in two important heterogeneous settings: first, when the test statistics are continuous but the p-values are weighted by some known independent weight vector, e.g., coming from co-data sets; second, when the test statistics are discretely distributed, as is the case for data representing frequencies or counts. Our new procedures are implemented in the R package `FDX`, see Junge and Döhler (2020).

AMS 2000 subject classifications: Primary 62H15; secondary 62Q05.
Keywords and phrases: False discovery exceedance, heterogeneous data, discrete hypothesis testing, weighted p-values, type I error rate control, step-down algorithm.

Received April 2020.

1. Introduction

1.1. Background

When many statistical tests are performed simultaneously, a ubiquitous way to account for the erroneous rejections of the procedure is the false discovery proportion (FDP), that is, the proportion of errors in the rejected sets, as introduced in the seminal paper Benjamini and Hochberg (1995). Most of the

related literature focuses on the expected value of this quantity, which is the false discovery rate (FDR), e.g., building procedures that improve the original Benjamini-Hochberg procedure by trying to adapt to some underlying structure of the data. In particular, a fruitful direction is to take into account the heterogeneous structure of the different tests. Heterogeneity may originate from various sources. The two main examples we have in mind, and which have been intensively investigated in the statistical literature recently, is heterogeneity caused by p-value weighting and discrete data.

Weighting p-values is a well-established and popular approach to improve the performance of multiple testing procedures. It can be traced back to Holm (1979) and that has been further developed specifically for FDR in, e.g., Genovese et al. (2006); Blanchard and Roquain (2008); Hu et al. (2010); Zhao and Zhang (2014); Ramdas et al. (2017). Here, the heterogeneity can be driven for instance by sample size, groups, or more generally by some covariates. In particular, finding optimal weighting in the sense of maximizing the number of true rejections has been investigated in Wasserman and Roeder (2006); Rubin et al. (2006); Roquain and van de Wiel (2009); Ignatiadis et al. (2016); Durand (2019). As a result, the weighted p-values have heterogeneous null distribution functions $\{F_i, 1 \leq i \leq m\}$ that must be properly taken into account by multiple testing procedures.

On the other hand, multiple testing for discrete distributions is a well identified research field (Tarone, 1990; Westfall and Wolfinger, 1997; Gilbert, 2005) that has received a growing attention in the last decade, see e.g., Heyse (2011); Heller and Gur (2011); Dickhaus et al. (2012); Habiger (2015); Chen et al. (2015); Döhler (2016); Chen et al. (2018); Döhler et al. (2018); Durand et al. (2019) and references therein. The most typical setting occurs when the data for each test is given by a contingency table. In that situation, the heterogeneity is induced by the fact that marginal counts naturally vary from one table to another. The approach is then to suitably combine the heterogeneous null distributions in order to compensate the natural conservativeness of individual discrete tests. Heyse (2011), for instance, uses the transform

$$\overline{F}(t) = m^{-1} \sum_{i=1}^{m} F_i(t), \quad t \in [0,1], \tag{1}$$

and applies the BH procedure to the transformed p-values $\{\overline{F}(p_i), 1 \leq i \leq m\}$. Unfortunately, this method does not rigorously control the FDR, as it has been proven in Döhler (2016); Döhler et al. (2018). Appropriate corrections of the $\overline{F}$ expression have been proposed in Döhler et al. (2018) in order to recover a rigorous FDR control.

1.2. FDX control

A common criticism of FDR is that it captures only the average behavior of the FDP. In particular, controlling the FDR does not prevent the FDP from

possessing undesirable fluctuations and we may aim at stochastically controlling the FDP in other ways. The most common alternative approach is to control the false probability exceedance (FDX) by requiring

$$\mathrm{FDX} = \mathbf{P}(\mathrm{FDP} > \alpha) \leq \zeta. \tag{2}$$

for $\alpha, \zeta \in (0,1)$. This corresponds to controlling the $(1-\zeta)$-quantile of the FDP distribution at level α, see, e.g., Genovese and Wasserman (2004); Perone Pacifico et al. (2004); Korn et al. (2004); Lehmann and Romano (2005); Genovese and Wasserman (2006); Romano and Wolf (2007); Guo et al. (2014); Delattre and Roquain (2015) and see Tan et al. (2019); Basu et al. (2020) for recent applications. Let us also mention that studying the probabilistic fluctuations of the FDP process is of interest in its own, see, e.g., Neuvial (2008); Roquain and Villers (2011); Delattre and Roquain (2011, 2016); Ditzhaus and Janssen (2019) and that controlling these fluctuations is used for various aims, as building FDP-confidence envelopes and post hoc bounds, see Genovese and Wasserman (2006); Goeman and Solari (2011); Hemerik et al. (2019); Katsevich and Ramdas (2020); Blanchard et al. (2020).

Among multiple testing procedures, step-down procedures have been shown to be particularly useful for FDX control. Two prominent step-down procedures have been proven to control the FDX under various distributional assumptions:

- The Lehmann-Romano procedure [LR], introduced in Lehmann and Romano (2005), is defined as the step-down procedure with critical values

$$\tau_\ell^{\mathrm{LR}} = \zeta \frac{\lfloor \alpha \ell \rfloor + 1}{m(\ell)} \ , 1 \leq \ell \leq m, \tag{3}$$

  where we denote

$$m(\ell) = m - \ell + \lfloor \alpha \ell \rfloor + 1. \tag{4}$$

  It has been shown to control the FDX under various dependence assumptions between the p-values, e.g., when each p-value under the null is independent of the family of the p-values under the alternative (Theorem 3.1 in Lehmann and Romano (2005)), which we will refer to (Indep0) below, or when the Simes inequality holds true among the family of true null p-values (Theorem 3.2 in Lehmann and Romano (2005)). Note that under the latter condition, it has also been proven later that the step-up version of [LR], that is, the step-up procedure using the critical values (3) also controls the FDX, see the proof of Theorem 3.1 in Guo et al. (2014).
- The procedure [LR] has been improved by the Guo-Romano procedure [GR], see Guo and Romano (2007), defined as the step-down procedure with critical values

$$\tau_\ell^{\mathrm{GR}} = \max\{t \in [0,1] \ : \ \mathbf{P}(\ \mathbf{Bin}[m(\ell), t] \geq \lfloor \alpha \ell \rfloor + 1) \leq \zeta\}, \ \ 1 \leq \ell \leq m, \tag{5}$$

  where $\mathbf{Bin}[n,p]$ denotes any random variable following a binomial distribution with parameters n and p. While rejecting more hypotheses than

[LR], the procedure [GR] controls the FDX under a stronger assumption: the null p-value family contains mutually independent variables and it is independent of the alternative p-value family, which we refer to (Indep) below.

1.3. Contributions

To our knowledge, FDX control under heterogeneity has been studied scarcely, see Genovese et al. (2006); Basu et al. (2020), and in some specific situations only. More precisely, Genovese et al. (2006) proposed such a control for weighted p-values via an augmentation type approach (see e.g., Dudoit and van der Laan, 2007), while the procedure in Basu et al. (2020) is developed in an empirical Bayes setting and thus corresponds to a Bayesian FDX control.

The global aim of our paper is to provide a general view on (frequentist) FDX control under heterogeneity by incorporating the null distribution functions $\{F_i, 1 \leq i \leq m\}$ of the p-values. To this end, we introduce suitable modifications of procedures [LR] and [GR]. More specifically, the contributions of this work are as follows:

- we introduce the heterogeneous Lehmann Romano procedure [HLR], which controls the FDX under (Indep0) and is a uniform improvement of [LR] (when the marginals of the null p-values are super-uniform, see (SuperUnif) further on);
- we introduce the heterogeneous Guo Romano procedure [HGR], which controls the FDX under (Indep) and is a uniform improvement of [GR] (under (SuperUnif));
- at the price of additional computational complexity, we introduce the Poisson-binomial procedure [PB], which controls the FDX under (Indep) and is a uniform improvement of [HLR] and [HGR];
- we apply this new technology to weighted p-values to provide the first weighted procedures that control the FDX (to our knowledge), called [wLR] and [wGR]. They are able to improve their non-weighted counterparts [LR] and [GR], respectively, see Section 4;
- in the discrete context, the new induced procedures [DLR], [DGR] are shown to be uniform improvements with respect to the continuous procedures [LR] and [GR], respectively. To the best of our knowledge, these are the first FDX controlling procedures tailored specifically to discrete p-value distributions. The improvement can be substantial, as we show both with simulated and real data examples, see Section 5.

The paper is organized as follows: Section 2 introduces the statistical setting, the procedures and FDX criterion, as well as a shortcut to compute our step-down procedures without evaluating the critical values. Section 3 is the main section of the paper, which introduces the new heterogeneous procedures and establishes their FDX controlling properties. Our methodology is then applied in two particular frameworks: new weighted procedures controlling the FDX

are derived in Section 4 while Section 5 is devoted to the case where the tests are discrete. Both sections include numerical illustrations. A discussion is provided in Section 6 and most of technical details are deferred to Appendix A. Appendix B gives additional numerical details for the simulations.

2. Framework

2.1. Setting

We use here a classical formal setting for heterogeneous nulls, see e.g., Döhler et al. (2018). We observe X, defined on an abstract probabilistic space, valued in an observation space $(\mathcal{X}, \mathfrak{X})$ and generated by a distribution P that belongs to a set $\mathcal{P}$ of possible distributions. We consider m null hypotheses for P, denoted $H_{0,i}$, $1 \leq i \leq m$, and we denote the corresponding set of true null hypotheses by $\mathcal{H}_0(P) = \{1 \leq i \leq m \ : \ H_{0,i} \text{ is satisfied by } P\}$. We also denote by $\mathcal{H}_1(P)$ the complement of $\mathcal{H}_0(P)$ in $\{1, \ldots, m\}$ and by $m_0(P) = |\mathcal{H}_0(P)|$ the number of true nulls.

We assume that there exists a set of p-values that is, a set of random variables $\{p_i(X), 1 \leq i \leq m\}$, valued in $[0,1]$. We introduce the following dependence assumptions between the p-values:

$$\text{for all } P \in \mathcal{P},\ \{p_i(X), i \in \mathcal{H}_0(P)\} \text{ is independent of } \{p_i(X), i \in \mathcal{H}_1(P)\}; \tag{Indep0}$$

$$\text{(Indep0) holds and for all } P \in \mathcal{P},\ \{p_i(X), i \in \mathcal{H}_0(P)\} \text{ consists of independent variables.} \tag{Indep}$$

Note that (Indep0) and (Indep) are both satisfied when all the p-values $p_i(X)$, $1 \leq i \leq m$, are mutually independent in the model $\mathcal{P}$. The (maximum) null cumulative distribution function of each p-value is denoted

$$F_i(t) = \sup_{P \in \mathcal{P} \,:\, i \in \mathcal{H}_0(P)} \{\mathbf{P}_{X \sim P}(p_i(X) \leq t)\}, \ \ t \in [0,1], \ \ 1 \leq i \leq m. \tag{6}$$

We let $\mathcal{F} = \{F_i, 1 \leq i \leq m\}$ that we assume to be *known* and we consider the following possible situations for the functions in $\mathcal{F}$:

$$\text{for all } i \in \{1, \ldots, m\},\ F_i \text{ is continuous on } [0,1] \tag{Cont}$$

$$\text{for all } i \in \{1, \ldots, m\}, \text{ there exists some finite set } \mathcal{A}_i \subset [0,1] \text{ such that } F_i \text{ is a step function, right continuous, that jumps only at some points of } \mathcal{A}_i. \tag{Discrete}$$

The case (Discrete) typically arises when for all $P \in \mathcal{P}$ and $i \in \{1, \ldots, m\}$, $\mathbf{P}_{X \sim P}(p_i(X) \in \mathcal{A}_i) = 1$ for some given finite sets $\mathcal{A}_i \subset [0,1]$. Throughout the paper, we will assume that we are either in the case (Cont) or (Discrete) and we denote $\mathcal{A} = \cup_{i=1}^m \mathcal{A}_i$, with by convention $\mathcal{A}_i = [0,1]$ when (Cont) holds. For

comparison with the homogeneous case, we will also make use of the following classical assumption:

$$\text{for all } i \in \{1,\dots,m\},\ F_i(t) \le t \text{ for all } t \in [0,1]. \tag{SuperUnif}$$

Remark 2.1. *Condition* (Discrete) *above requires that the sets* $\mathcal{A}_i$ *are finite. While this condition is met, e.g., when Fisher's exact tests are performed, it is not satisfied for other classical discrete tests, as the one-sided Poisson test for instance. Interestingly, we can also deal with such cases, because all our results remain valid when the sets* $\mathcal{A}_i$ *are countable with zero as only possible accumulation point.*

2.2. False discovery exceedance and step-down procedures

In general, a multiple testing procedure is defined as a random subset $R = R(X) \subset \{1,\dots,m\}$ which corresponds to the indices of the rejected nulls. For $\alpha \in (0,1)$, the false discovery exceedance of R is defined as follows:

$$\mathrm{FDX}_\alpha(R,P) = \mathbf{P}_{X\sim P}\left(\frac{|R(X)\cap \mathcal{H}_0(P)|}{|R(X)|\vee 1} > \alpha\right),\quad P\in\mathcal{P}. \tag{7}$$

In this paper, we consider particular multiple testing procedures, called step-down procedures. Given some p-value family $(p_i)_{1\le i\le m}$ and some non-decreasing sequence $(\tau_\ell)_{1\le \ell\le m} \in [0,1]^m$, the step-down procedure with critical values $(\tau_\ell)_{1\le\ell\le m}\in[0,1]^m$ rejects the null hypotheses corresponding to the set

$$R = \{i\in\{1,\dots,m\} : p_i(X) \le \tau_{\widehat{\ell}}\} \qquad \text{where} \tag{8}$$

$$\widehat{\ell} = \max\{\ell\in\{0,\dots,m\} : \forall \ell' \le \ell,\ p_{\sigma(\ell')} \le \tau_{\ell'}\},\quad (\text{convention } p_{\sigma(0)}=0), \tag{9}$$

for which $p_{\sigma(1)} \le \cdots \le p_{\sigma(m)}$ denotes the p-values $\{p_i(X), 1\le i\le m\}$ ordered increasingly (for some data-dependent permutation σ).

2.3. Transformation function family and computational shortcut

In this paper, the critical values will be obtained by inverting specific functionals, that is,

$$\tau_\ell = \xi_\ell^{-1}(\zeta) = \max\{t\in\mathcal{A} : \xi_\ell(t)\le\zeta\}, (\tau_\ell = 0 \text{ if the set is empty}),\ \ 1\le\ell\le m, \tag{10}$$

for $\xi_\ell : [0,1]\mapsto[0,\infty)$, $1\le\ell\le m$, a given set of functions. In order for (10) to be well-defined and the corresponding $\ell\mapsto\tau_\ell$ to be non-decreasing, we will say that the function set $\{\xi_\ell, 1\le\ell\le m\}$ is a *transformation function family* if it satisfies the following conditions:

$$\begin{aligned}&\text{for all } \ell\in\{1,\dots,m\},\ \xi_\ell \text{ is a non-decreasing function;}\\ &\text{for all } t\in[0,1] \text{ and all } \ell\in\{1,\dots,m-1\}, \text{ we have } \xi_{\ell+1}(t)\le\xi_\ell(t);\\ &\text{in case (Cont), for all } \ell\in\{1,\dots,m\},\ \xi_\ell \text{ is continuous on } [0,1].\end{aligned} \tag{11}$$

For instance, the critical values of the procedure [LR] can be rewritten as (10) for the functions

$$\xi_\ell^{\mathrm{LR}}(t) = \frac{m(\ell)}{\lfloor \alpha\ell \rfloor + 1} t, \;\; t \in [0,1], \;\; 1 \le \ell \le m. \tag{12}$$

We easily check that the function set $\{\xi_\ell^{\mathrm{LR}}, 1 \le \ell \le m\}$ is a family of transformation functions (in the sense of (11)). Indeed, $\frac{m-\ell+i}{i}$ is non-increasing both in $\ell \in \{1,\dots,m\}$ and $i \in \{1,\dots,\lfloor \alpha m \rfloor + 1\}$. A second example is given by the procedure [GR] for which

$$\xi_\ell^{\mathrm{GR}}(t) = \mathbf{P}(\mathbf{Bin}[m(\ell), t] \ge \lfloor \alpha\ell \rfloor + 1), \quad 1 \le \ell \le m, \; t \in [0,1], \tag{13}$$

can be proved to form a family of transformation functions. Indeed, the only non-obvious argument to prove (11) is that for a fixed $t \in [0,1]$, and $\ell \in \{1,\dots,m-1\}$ we have $\xi_{\ell+1}^{\mathrm{GR}}(t) \le \xi_\ell^{\mathrm{GR}}(t)$. This is due to the fact that $\mathbf{P}(\mathbf{Bin}[m-\ell+i, t] \ge i) = \mathbf{P}(\mathbf{Bin}[m-\ell+i, 1-t] \le m-\ell)$ is non-increasing both in i and ℓ.

Finally, because of the inversion, computing the critical values via (10) can be time consuming. Fortunately, computing the critical values is actually not necessary if we are solely interested in determining the rejection set R given by (8). As the following result shows, we can determine R by working directly with the transformation functions.

Proposition 2.1. *Let us consider any transformation function family $\{\xi_\ell, 1 \le \ell \le m\}$ and the corresponding critical values τ_ℓ, $1 \le \ell \le m$, defined by (10). Then, for all $P \in \mathcal{P}$, with P-probability 1, the step-down procedure R with critical values $(\tau_\ell)_{1 \le \ell \le m}$ can equivalently be written as*

$$R = \{i \in \{1,\dots,m\} \,:\, \tilde{p}_i \le \zeta\}; \tag{14}$$

$$\tilde{p}_i = \max_{\substack{1 \le \ell \le m \\ p_{\sigma(\ell)} \le p_i}} \{\xi_\ell(p_{\sigma(\ell)})\}, \;\; 1 \le i \le m. \tag{15}$$

Proposition 2.1 is proved in Appendix A.2.

3. New FDX controlling procedures

In this section, we introduce new procedures R that control the false discovery exceedance at some level $\zeta \in (0,1)$, that is,

$$\text{for all } P \in \mathcal{P}, \; \mathrm{FDX}_\alpha(R, P) \le \zeta, \tag{16}$$

while incorporating the family $\{F_i, 1 \le i \le m\}$ in an appropriate way.

3.1. Tool

Our main mathematical tool is the following bound: For any step-down procedure R with critical values $\tau = (\tau_\ell)_{1 \le \ell \le m}$, we have

$$\sup_{P \in \mathcal{P}} \{\mathrm{FDX}_\alpha(R, P)\} \le B(\tau, \alpha) \tag{17}$$

$$\text{for } B(\tau,\alpha) = \sup_{1\le\ell\le m} \sup_{\substack{P\in\mathcal{P}\\ |\mathcal{H}_0(P)|\le m(\ell)}} \mathbf{P}_{X\sim P}\left(\sum_{i\in\mathcal{H}_0(P)} \mathbb{1}\{p_i(X)\le\tau_\ell\} \ge \lfloor\alpha\ell\rfloor + 1\right). \tag{18}$$

Inequality (17) is valid under the distributional assumption (Indep0). This bound comes from a reformulation of Theorem 5.2 in Roquain (2011) in our heterogenous framework, see Theorem A.1 in Appendix A below. Our new procedures are derived by further upper-bounding $B(\tau,\alpha)$ via various probabilistic devices. More specifically, we will introduce several transformation function families $\{\xi_\ell, 1\le\ell\le m\}$ such that for all $\tau=\{\tau_\ell\}_\ell$,

$$B(\tau,\alpha) \le \sup_{1\le\ell\le m}\{\xi_\ell(\tau_\ell)\}.$$

According to (17), the step-down procedure using the corresponding critical values (10) will then control the FDX in the sense of (16).

3.2. Heterogeneous Lehmann-Romano procedure

By using the Markov inequality, we obtain

$$B(\tau,\alpha) \le \sup_{1\le\ell\le m} \sup_{\substack{P\in\mathcal{P}\\ |\mathcal{H}_0(P)|\le m(\ell)}} \frac{\sum_{i\in\mathcal{H}_0(P)} F_i(\tau_\ell)}{\lfloor\alpha\ell\rfloor+1} = \sup_{1\le\ell\le m} \frac{\sum_{j=1}^{m(\ell)} F_{(j)}(\tau_\ell)}{\lfloor\alpha\ell\rfloor+1}, \tag{19}$$

where $F_{(1)}(t)\ge\cdots\ge F_{(m)}(t)$ denotes the values of $\{F_i(t), 1\le i\le m\}$ ordered decreasingly. Bounding the above quantity by ζ entails the following procedure.

Definition 3.1. *The heterogeneous Lehmann-Romano procedure, denoted by* [HLR]*, is defined as the step-down procedure using the critical values defined by*

$$\tau_\ell^{HLR} = \max\{t\in\mathcal{A} : \xi_\ell^{HLR}(t)\le\zeta\}, \quad 1\le\ell\le m; \tag{20}$$

$$\xi_\ell^{HLR}(t) = \frac{\sum_{j=1}^{m(\ell)} F_{(j)}(t)}{\lfloor\alpha\ell\rfloor+1}, \quad 1\le\ell\le m,\ t\in[0,1], \tag{21}$$

where $F_{(1)}(t)\ge\cdots\ge F_{(m)}(t)$ denotes the values of $\{F_i(t), 1\le i\le m\}$ ordered decreasingly and $m(\ell)$ is defined by (4).

The quantity $\xi_\ell^{\mathrm{HLR}}(t)$ is thus similar to $\xi_\ell^{\mathrm{LR}}(t)$, in which t has been replaced by the average of the $m(\ell)$ largest values of $\{F_i(t), 1\le i\le m\}$. To check that the set $\{\xi_\ell^{\mathrm{HLR}}, 1\le\ell\le m\}$ is a transformation function family in the sense of (11), we note that $\frac{1}{m(\ell)}\sum_{j=1}^{m(\ell)} F_{(j)}(t)$ is non-increasing in ℓ (averaging on smaller values makes the average smaller) and continuous in t under (Cont) (because $t\mapsto(F_i(t))_{1\le i\le m}$ is continuous in that case and $x\in(\mathbb{R}^m, \|\cdot\|_\infty)\mapsto N^{-1}\sum_{k=1}^N x_{(k)}\in(\mathbb{R},|\cdot|)$ is 1-Lipschitz for any $N\in\{1,\dots,m\}$, see Lemma A.2).

In the classical case (SuperUnif), we have $\xi_\ell^{\mathrm{HLR}}(t)\le\xi_\ell^{\mathrm{LR}}(t)$ for all $t\in[0,1]$ and $1\le\ell\le m$. Hence, [HLR] is less conservative than [LR] in that situation.

A technical detail is that this only holds almost surely because the range $\mathcal{A}$ in (20) can be different from $[0,1]$ in the case (Discrete). We have established the following result.

Proposition 3.1. *In the setting defined in Section 2.1, the procedure* [HLR] *satisfies the following*

- *Under* (Indep0)*,* [HLR] *controls the* FDX *in the sense* (16)*;*
- *Under* (SuperUnif)*, the set of nulls rejected by* [HLR] *contains the one of* [LR] *with P-probability* 1*, for all $P \in \mathcal{P}$.*

3.3. Poisson-binomial procedure

Here, we propose to bound (18) by using the Poisson-binomial distribution. To this end, recall that the Poisson-Binomial distribution of parameters $\pi = (\pi_i)_{1\le i\le n} \in [0,1]^n$, denoted $\mathbf{PBin}[\pi]$ below, corresponds to the distribution of $\sum_{i=1}^n \varepsilon_i$, where the ε_i are all independent and each ε_i follows a Bernoulli distribution of parameter π_i for $1 \le i \le n$.

First note that for all $i \in \mathcal{H}_0(P)$ and $t \in [0,1]$, we have that $\mathbb{1}\{p_i(X) \le t\}$ is stochastically upper bounded by a Bernoulli variable of parameter $F_i(t)$, see (6). As a consequence, by assuming (Indep), we have for all critical values $(\tau_\ell)_{1\le\ell\le m}$,

$$\begin{aligned} B(\tau,\alpha) &\le \sup_{1\le\ell\le m} \sup_{\substack{A\subset\{1,\dots,m\}\\ |A|\le m(\ell)}} \mathbf{P}\left(\mathbf{PBin}\left[(F_i(\tau_\ell))_{i\in A}\right] \ge \lfloor\alpha\ell\rfloor + 1\right) \\ &= \sup_{1\le\ell\le m} \mathbf{P}\left(\mathbf{PBin}\left[(F_{(j)}(\tau_\ell))_{1\le j\le m(\ell)}\right] \ge \lfloor\alpha\ell\rfloor + 1\right). \end{aligned} \tag{22}$$

Bounding the latter by ζ leads to the following procedure.

Definition 3.2. *The Poisson-binomial procedure, denoted by* [PB]*, is defined as the step-down procedure using the critical values*

$$\tau_\ell^{PB} = \max\{t \in \mathcal{A} : \xi_\ell^{PB}(t) \le \zeta\}, \quad 1 \le \ell \le m; \tag{23}$$

$$\xi_\ell^{PB}(t) = \mathbf{P}\left(\mathbf{PBin}\left[(F_{(j)}(t))_{1\le j\le m(\ell)}\right] \ge \lfloor\alpha\ell\rfloor + 1\right), \quad 1 \le \ell \le m,\ t \in [0,1], \tag{24}$$

where $F_{(1)}(t) \ge \cdots \ge F_{(m)}(t)$ denotes the values of $\{F_i(t), 1 \le i \le m\}$ ordered decreasingly and $m(\ell)$ is defined by (4)*.*

Let us now check that $\{\xi_\ell^{\mathrm{PB}}, 1 \le \ell \le m\}$ is a transformation function family, that is, it satisfies (11). The continuity assumption holds because, under (Cont), the mapping $t \in [0,1] \mapsto (F_{(j)}(t))_{1\le j\le m(\ell)}$ is continuous (argument similar to above) and the cumulative distribution function of $\mathbf{PBin}[\pi]$ is a continuous function of $\pi \in [0,1]^n$. The monotonic property $\xi_{\ell+1}^{\mathrm{HGR}}(t) \le \xi_\ell^{\mathrm{HGR}}(t)$ comes from the fact that the probability $\mathbf{P}(\mathbf{PBin}\left[(F_{(j)}(t))_{1\le j\le m-\ell+i}\right] \ge i)$ is equal to $\mathbf{P}(\mathbf{PBin}\left[(1-F_{(j)}(t))_{1\le j\le m-\ell+i}\right] \le m-\ell)$, which is non-increasing both in i and ℓ.

In addition, under (SuperUnif), the distribution $\mathbf{PBin}\left[(F_{(j)}(t))_{1\leq j\leq m(\ell)}\right]$ is stochastically smaller than the distribution $\mathbf{Bin}[m(\ell),t]$. We have proved the following holds.

Proposition 3.2. *In the setting defined in Section 2.1, the procedure* [PB] *satisfies the following*

- *Under* (Indep), [PB] *controls the* FDX *in the sense* (16)*;*
- *Under* (SuperUnif)*, the set of nulls rejected by* [PB] *contains the one of* [GR] *with P-probability* 1*, for all $P\in\mathcal{P}$.*

Note that the procedure [PB] relies on Poisson-binomial tail probabilities, see (24), which is computationally more demanding than using the binomial distribution. Nevertheless, algorithms are available and implemented in software for determining these probabilities (see Junge (2020) and the references given there). The critical values in (23) can subsequently be determined by using a standard numerical root finding algorithm (for details, see (Junge and Döhler, 2020)).

To avoid determining Poisson-binomial tail probabilities, we can use a slightly conservative approach, only relying on the binomial distribution, which we describe in the next section.

3.4. Heterogeneous Guo-Romano procedure

In this section, we further upper-bound (25) by using that any $\mathbf{PBin}\left[(\pi_i)_{1\leq i\leq n}\right]$ variable is stochastically upper-bounded by a $\mathbf{Bin}\left[n,1-\left(\prod_{i=1}^n(1-\pi_i)\right)^{1/n}\right]$ variable (see Example 1.A.25 in Shaked, M. and Shanthikumar, J.G., 2007). This yields

$$B(\tau,\alpha)\leq\sup_{1\leq\ell\leq m}\mathbf{P}\left(\mathbf{Bin}\left[m(\ell),\tilde{F}_{m(\ell)}(\tau_\ell)\right]\geq\lfloor\alpha\ell\rfloor+1\right),\tag{25}$$

where we let

$$\tilde{F}_j(t)=1-\left(\prod_{j'=1}^{j}(1-F_{(j')}(t))\right)^{1/j},\quad 1\leq j\leq m,\ t\in[0,1],\tag{26}$$

where $F_{(1)}(t)\geq\cdots\geq F_{(m)}(t)$ denotes the values of $\{F_i(t),1\leq i\leq m\}$ ordered decreasingly.

This reasoning suggests another heterogeneous procedure, based on the binomial distribution. Since [GR] also uses the binomial device, we name this new procedure the heterogeneous Guo-Romano procedure.

Definition 3.3. *The heterogeneous Guo-Romano procedure, denoted by* [HGR]*, is defined as the step-down procedure using the critical values defined by*

$$\tau_\ell^{HGR}=\max\{t\in\mathcal{A}\,:\,\xi_\ell^{HGR}(t)\leq\zeta\},\quad 1\leq\ell\leq m;\tag{27}$$

$$\xi_\ell^{HGR}(t) = \mathbf{P}\left(\mathbf{Bin}\left[m(\ell), \tilde{F}_{m(\ell)}(t)\right] \geq \lfloor \alpha\ell \rfloor + 1\right), \quad 1 \leq \ell \leq m,\ t \in [0,1], \quad (28)$$

where $\tilde{F}_j(t)$ *is defined in* (26) *and* $m(\ell)$ *is defined by* (4).

The condition (11) also holds in that case. However, the proof of monotonicity of $\xi_\ell^{\mathrm{HGR}}(t)$ is slightly more involved than above and is deferred to Lemma A.1. In addition, since under (SuperUnif) we have $\tilde{F}_{m(\ell)}(t) \leq t$, we deduce that [HGR], although more conservative than [PB], is still an uniform improvement over [GR].

Proposition 3.3. *In the setting defined in Section 2.1, the procedure* [HGR] *satisfies the following*

- *Under* (Indep), [HGR] *controls the* FDX *in the sense* (16)*;*
- *Under* (SuperUnif), *the set of nulls rejected by* [HGR] *contains the one of* [GR] *with* P*-probability* 1*, for all* $P \in \mathcal{P}$.

Remark 3.1. *The numerical results in Sections 4 and 5 suggest that the conservatism of* [HGR] *with respect to* [PB] *is usually quite small. In addition, since the computational effort required by* [HGR] *is comparable to that of* [GR]*, the gain in efficiency may be great, especially for large* m*. We therefore think that* [HGR] *may be especially useful for very high dimensional heterogeneous data.*

Remark 3.2. *We can also define a non-adaptive version of* [HGR]*, defined as the step-down procedure of critical values* (10) *based on the transformation functional*

$$\xi_\ell(t) = \mathbf{P}\left(\mathbf{Bin}\left[m, \tilde{F}(t)\right] \geq \lfloor \alpha\ell \rfloor + 1\right), \quad 1 \leq \ell \leq m,\ t \in [0,1],$$

where $\tilde{F}(t) = 1 - \left(\prod_{j=1}^m (1 - F_j(t))\right)^{1/m}$*. While being more conservative than* [HGR]*, it still controls the* FDX *in the sense* (16)*. Hence, while controlling the FDR for heterogeneous tests is linked to the arithmetic average of the* F_i*'s (see Section 1.1) and requires some additional modifications (see Döhler et al., 2018), our results show that FDX control is linked to simple geometric averaging without the need of any additional correction.*

4. Application to weighting

It is well known that p-value weighting can improve the power of multiple testing procedures, see e.g., Genovese et al. (2006); Roquain and van de Wiel (2009); Ignatiadis et al. (2016); Durand (2019); Ramdas et al. (2017) and references therein. However, to the best of our our knowledge, except for the augmentation approach described in Genovese et al. (2006), no methods are available that incorporate weighting for FDX control. We show in this section that such methods can be obtained directly from the bounds on $B(\tau, \alpha)$ introduced in Section 3.

Throughout this section, we consider the standard setting for which we have at hand a p-value family satisfying (Cont) *and* (SuperUnif). As explained in our introduction section (see references therein), while the null distributions of the p-values are typically uniform, the point is that they can have heterogeneous alternative distributions, so that it could be desirable to weigh the p-values in some way. For this, we consider a fixed weight vector $(w_i)_{1\le i\le m} \in \mathbb{R}_+^m$. The ordered weights are denoted $w_{(1)} \ge w_{(2)} \ge \cdots \ge w_{(m)}$, the average weight is denoted $\overline{w} = m^{-1}\sum_{i=1}^m w_i$ and the average over the j largest weights is denoted by $\overline{w}_j = j^{-1}\sum_{j'=1}^j w_{(j')}$.

Since the heterogeneous procedures [HLR], [PB] and [HGR] introduced in Section 3 yield valid control for any collection of distribution functions $\{F_i, 1 \le i \le m\}$, it is possible to use very flexible weighting schemes. In order to limit the scope of this paper, we consider only two simple types of weighting approaches in more detail:

- for *arithmetic mean weighting* (abbreviated in what follows as AM), we define the weighted p-value family as
$$p_i^w = p_i\,\bar{w}/w_i, \quad 1 \le i \le m. \tag{29}$$
The weighted p-values thus have the heterogeneous distribution functions
$$F_i^{\mathrm{AM}}(t) = \left(\frac{w_i}{\bar{w}}t\right) \wedge 1, \quad t \in [0,1], \quad 1 \le i \le m, \tag{30}$$
under the null. This corresponds to classical weighting approaches established in the multiple testing literature for various criteria, like FWER and FDR control.
- for *geometric mean weighting* (abbreviated as GM), we define
$$p_i^w = 1 - (1-p_i)^{\bar{w}/w_i}, \quad 1 \le i \le m. \tag{31}$$
The weighted p-values therefore have the following heterogeneous distribution functions under the null:
$$F_i^{\mathrm{GM}}(t) = 1 - (1-t)^{w_i/\bar{w}}, \quad t \in [0,1], \quad 1 \le i \le m. \tag{32}$$

Thus, combining these two weighting approaches with the three heterogeneous procedures introduced in the previous section yields a total of six weighted procedures which we discuss in more detail below. Note that a Taylor expansion yields $F_i^{\mathrm{AM}}(t) \approx F_i^{\mathrm{GM}}(t)$ for small values of t. Therefore, we expect that AM and GM procedures will yield similar rejection sets for small p-values.

4.1. Weighted Lehmann-Romano procedures

Applying the strategy of Section 3.2 with the c.d.f. set $\{F_i^{\mathrm{AM}}, 1 \le i \le m\}$, we can use the transformation function family given by
$$\xi_\ell^{\text{wLR-AM}}(t) = \frac{m(\ell)}{\lfloor \alpha\ell \rfloor + 1} \times \frac{\overline{w}_{m(\ell)}}{\bar{w}} t, \quad 1 \le \ell \le m,\ t \in [0,1].$$

This gives rise to an FDX controlling procedure, which we call the AM-weighted Lehmann-Romano procedure, denoted by [wLR-AM]. It is defined as the step-down procedure using the weighted p-values (29) and the critical values

$$\tau_\ell^{\text{wLR-AM}} = \zeta \frac{\lfloor \alpha \ell \rfloor + 1}{\sum_{j=1}^{m(\ell)} w_{(j)}} \bar{w} = \tau_\ell^{\text{LR}} \cdot \frac{\bar{w}}{\overline{w}_{m(\ell)}}, \qquad 1 \le \ell \le m.$$

In particular, if the weight vector is uniform, that is, $w_i = 1$ for all i, then [wLR-AM] reduces to [LR].

Similarly to above, applying the strategy of Section 3.2 with the c.d.f. set $\{F_i^{\text{GM}}, 1 \le i \le m\}$ yields the transformation function family

$$\xi_\ell^{\text{wLR-GM}}(t) = \frac{1}{\lfloor \alpha \ell \rfloor + 1} \sum_{j=1}^{m(\ell)} (1 - (1-t)^{w_{(j)}/\bar{w}}), \quad 1 \le \ell \le m,\ t \in [0,1].$$

This gives rise to an FDX controlling procedure, which we call the GM-weighted Lehmann-Romano procedure, denoted by [wLR-GM]. It is defined as the step-down procedure using the weighted p-values (31) and the critical values

$$\tau_\ell^{\text{wLR-GM}} = \max\{t \in [0,1] : \xi_\ell^{\text{wLR-GM}}(t) \le \zeta\}, \quad 1 \le \ell \le m.$$

In general, no domination relationship holds between the procedures [wLR-GM] and [wLR-AM]. Finally, again, in case of uniform weighting, [wLR-GM] reduces to [LR].

4.2. Weighted Poisson-binomial procedures

Applying the strategy of Section 3.3 with the c.d.f. sets $\{F_i^{\text{AM}}, 1 \le i \le m\}$ and $\{F_i^{\text{GM}}, 1 \le i \le m\}$, we can use the two transformation function families given by

$$\xi_\ell^{\text{wPB-AM}}(t) = \mathbf{P}\left(\mathbf{PBin}\left[\left(\left(\frac{w_{(j)}}{\bar{w}} t\right) \wedge 1\right)_{1 \le j \le m(\ell)}\right] \ge \lfloor \alpha \ell \rfloor + 1\right);$$
$$\xi_\ell^{\text{wPB-GM}}(t) = \mathbf{P}\left(\mathbf{PBin}\left[(1 - (1-t)^{w_{(j)}/\bar{w}})_{1 \le j \le m(\ell)}\right] \ge \lfloor \alpha \ell \rfloor + 1\right),$$

for $1 \le \ell \le m, t \in [0,1]$, to define new step-down procedures, denoted [wPB-AM] and [wPB-GM] respectively, that both ensure FDX control.

4.3. Weighted Guo-Romano procedures

We apply here the strategy of Section 3.4 for the c.d.f. sets $\{F_i^{\text{AM}}, 1 \le i \le m\}$ and $\{F_i^{\text{GM}}, 1 \le i \le m\}$. According to (26), let us define

$$\tilde{F}_j^{\text{AM}}(t) = 1 - \left(\prod_{j'=1}^{j} \left(1 - \left(\frac{w_{(j')}}{\bar{w}} t\right) \wedge 1\right)\right)^{1/j};$$

$$\tilde{F}_j^{\mathrm{GM}}(t) = 1 - \left(\prod_{j'=1}^{j} (1-t)^{w_{(j')}/\bar{w}} \right)^{1/j} = 1 - (1-t)^{\overline{w}_j/\bar{w}},$$

for $1 \le j \le m$, $t \in [0,1]$. This gives rise to the transformation function families

$$\xi_\ell^{\text{wGR-AM}}(t) = \mathbf{P}\left(\mathbf{Bin}\left[m(\ell), \tilde{F}_{m(\ell)}^{\mathrm{AM}}(t) \right] \ge \lfloor \alpha \ell \rfloor + 1 \right), \quad 1 \le \ell \le m,\ t \in [0,1];$$
$$\xi_\ell^{\text{wGR-GM}}(t) = \mathbf{P}\left(\mathbf{Bin}\left[m(\ell), \tilde{F}_{m(\ell)}^{\mathrm{GM}}(t) \right] \ge \lfloor \alpha \ell \rfloor + 1 \right), \quad 1 \le \ell \le m,\ t \in [0,1].$$

Critical values $\tau^{\text{wGR-AM}}$ and $\tau^{\text{wGR-GM}}$ are obtained via (10) from families $\xi^{\text{wGR-AM}}$ and $\xi^{\text{wGR-GM}}$, respectively. This yields two new FDX controlling step-down procedures that are denoted by [wGR-AM] and [wGR-GM], respectively. Note that, similar to arithmetic weighting for the [LR] procedure, geometric weighting leads to a simple transformation of the original [GR] critical values, given by

$$\tau_\ell^{\text{wGR-GM}} = 1 - (1 - \tau_\ell^{\mathrm{GR}})^{\bar{w}/\overline{w}_{m(\ell)}}. \tag{33}$$

Thus, this particular procedure combines simplicity with a close relationship to the original Guo-Romano procedure. By contrast, as for the heterogeneous version, the weighted Poisson-binomial procedures require the evaluation of the Poisson-binomial distribution function which may be computationally demanding for large m. The weighted Guo-Romano procedures, on the other hand, while possibly sacrificing some power, only require evaluation of the standard binomial distribution.

4.4. Analysis of RNA-Seq data

We revisit an analysis of the RNA-Seq data set 'airway' using results from the independent hypothesis weighting (IHW) approach (for details, see Ignatiadis et al., 2016 and the vignette accompanying its software implementation). Loosely speaking, this method aims to increase power by assigning a weight w_i to each hypothesis and subsequently applying e.g., the Bonferroni or the Benjamini-Hochberg procedure [BH] to the weighted p-values while aiming for control of FWER or FDR.

In what follows, we present some results for weighted FDX control, using the procedures introduced in Sections 4.2 and 4.3. For this data set we have $m = 64102$ and the weights $w_1, \ldots, w_m$ are taken from the output of the ihw function from the bioconductor package 'IHW'. For the sake of illustration we assume the p-values to be independent. A large portion (about 45%) of these weights are 0, Figure 1 presents a histogram of the (strictly) positive weights.

Table 1 shows that controlling the mean (i.e. FDR) or the median of the FDP leads to a similar number of rejections.

For both error rates, incorporating weights leads to similar gains in power. For weighted FDX control, the more conservative weighted Guo-Romano procedures exhibit only a slight loss of power with respect to the weighted Poisson-binomial

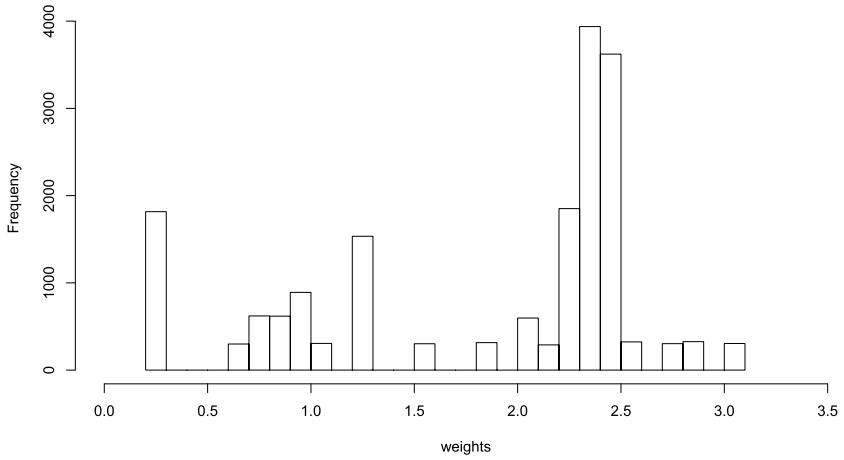

FIG 1. *Histogram of positive weights generated by the ihw function for the airway data.*

approaches. The difference between arithmetic and geometric weighting is negligible for this data.

Figure 2 indicates that for the FDX controlling procedures, the mapping of the confidence level to the number of rejections is quite flat. This means that statements about the FDP can be made with high confidence without losing too much power. For instance, requiring that FDP $\leq 10\%$ with confidence at least 95% still allows for 4145 and 4771 rejections using the [GR] and [wGR-GM] procedures, respectively.

TABLE 1
Number of rejections for the airway data. The FDR procedures control FDR *at level* 10%*, the FDX procedures control* $P(\text{FDP} > 10\%) \leq 0.5$.

	[BH]	[wBH]	[GR]	[wPB-AM]	[wPB-GM]	[wGR-AM]	[wGR-GM]
Rejections	4099	4896	4243	4868	4865	4853	4852

5. Application to discrete tests

5.1. Discrete FDX procedures

Discrete FDX procedures can be defined in a straightforward way by directly using the distribution functions $F_1, \ldots, F_m$ of the discretely distributed p-values. The prototypical example we have in mind are multiple conditional tests like Fisher's exact test. In this case, discreteness and heterogeneity arise from conditioning on the observed table margins. We denote the resulting heterogeneous procedures from section 3 by [DLR] (for [HLR]), [DPB] (for [HPB]) and [DGR] (for [HGR]).

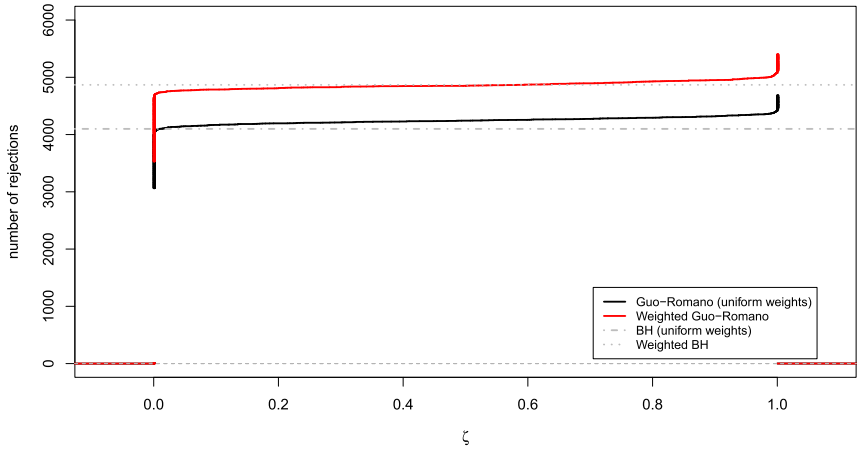

FIG 2. *Number of rejections (y-axis) for the airway data when using the* [GR] *and* [wGR-GM] *procedure. Both procedures control the tail probabilities (on the x-axis) for the* FDP *exceeding* 10%. *The horizontal lines represent the rejections of the BH and weighted BH procedures at* FDR-*level* 10%.

5.2. Simulation study

We now investigate the power of the [DLR], [DPB] and [DGR] procedures in a simulation study similar to those described in Gilbert (2005), Heller and Gur (2011) and Döhler et al. (2018). We focus on comparing the performance of the new discrete procedures to their continuous counterparts. Since the analysis with [DPB] is computationally demanding, we are also interested in investigating the performance of the slightly more conservative, but numerically more efficient [DGR] procedure. Finally, as above, we also include [BH] (Benjamini-Hochberg procedure) as a benchmark.

5.2.1. Simulated scenarios

We simulate a two-sample problem in which a vector of m independent binary responses ("adverse events") is observed for each subject in two groups, where each group consists of $N = 25$ subjects. Then, the goal is to simultaneously test the m null hypotheses H_{0i} : "$p_{1i} = p_{2i}$", $i = 1, \ldots, m$, where p_{1i} and p_{2i} are the success probabilities for the ith binary response in group 1 and 2, respectively. Before we describe the simulation framework in more detail, we explain how this set-up leads to discrete and heterogeneous p-value distributions. Suppose we have simulated two vectors of dimension m where each component represents a count in $\{0, \ldots, 25\}$. This data can be represented by m contingency tables. Now each hypothesis is tested using Fisher's exact test (two-sided) for each contingency table, which is performed by conditioning on the (simulated) pair of marginal counts. Thus, we can determine for every contingency table i the

discrete distribution function F_i of the p-values for Fisher's exact test under the null hypothesis. For differing (simulated) contingency tables, these induced distributions will generally be heterogeneous and our inference is conditionally on the marginal counts.

We take $m = 800, 2000$ where $m = m_1 + m_2 + m_3$ and data are generated so that the response is $Bernoulli(0.01)$ at m_1 positions for both groups, $Bernoulli(0.10)$ at m_2 positions for both groups and $Bernoulli(0.10)$ at m_3 positions for group 1 and $Bernoulli(q)$ at m_3 positions for group 2 where $q = 0.15, 0.25, 0.4$ represents weak, moderate and strong effects, respectively. The null hypothesis is true for the m_1 and m_2 positions while the alternative hypothesis is true for the m_3 positions. We also take different configurations for the proportion of false null hypotheses, m_3 is set to be 10%, 30% and 80% of the value of m, which represents small, intermediate and large proportion of effects, respectively (the proportion of true nulls π_0 is 0.9, 0.7, 0.2, respectively). Then, m_1 is set to be 20%, 50% and 80% of the number of true nulls (that is, $m - m_3$) and m_2 is taken accordingly as $m - m_1 - m_3$.

For each of the 54 possible parameter configurations specified by m, m_3, m_1 and q, 10000 Monte Carlo trials are performed, that is, 10000 data sets are generated and for each data set, an unadjusted two-sided p-value from Fisher's exact test is computed for each of the m positions, and the multiple testing procedures mentioned above are applied at level $\alpha = 0.05$ and $\zeta = 0.5$. The power of each procedure was estimated as the fraction of the m_3 false null hypotheses that were rejected, averaged over these 10000 simulations (TDP, true discovery proportion). Note that while our procedures are designed to control the FDP conditionally on the marginal counts, our power results are presented in an unconditional way for the sake of simplicity. For random number generation the R-function *rbinom* was used. The two-sided p-values from Fisher's exact test were computed using the R-function *fisher.test*.

5.2.2. Results

Table 3 in Appendix B shows that the (average) power of the compared procedures depends primarily on the strength of the signal $q_3 \in \{0.15, 0.25, 0.4\}$. More specifically, Figure 3 contains some typical plots of the simulation results.

- For $q_3 = 0.15$, the power of [BH], [LR] and [GR] is practically zero, whereas the discrete procedures are able to reject at least a few hypotheses, see panel (a) of Figure 3.
- For $q_3 = 0.25$, the power of [BH] and [LR] stays close to zero, [GR] performs slighty better and the discrete variants perform best as illustrated in panel (b) of Figure 3.
- For $q_3 = 0.4$, the power of [LR] stays close to zero, while [BH] now rejects a significant amount of hypotheses. The [DPB] and [DGR] procedures perform best. If there is a large amount of alternatives, [GR] performs better than [DLR] (see panel (c) of Figure 3). In the other cases, [GR] is outperformed by [DLR] (see panel (d) of Figure 3).

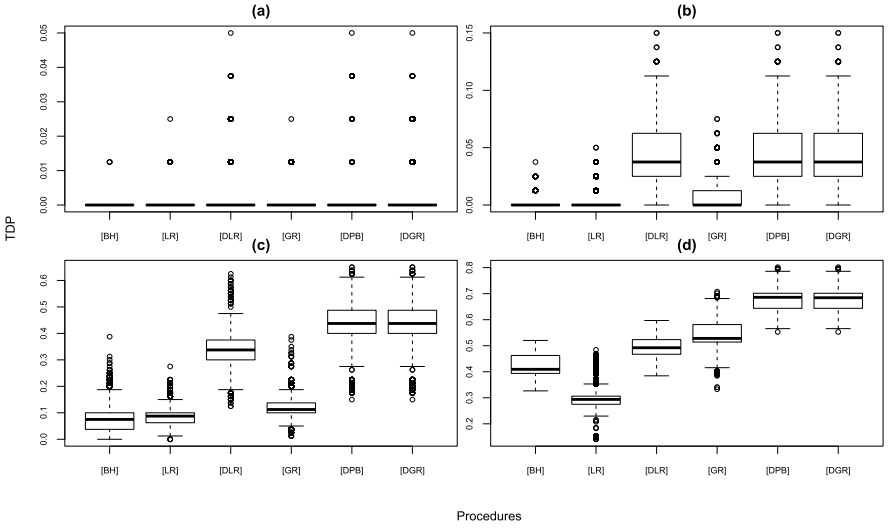

FIG 3. *Boxplots of the simulated true discovery proportions (TDP) for the* [LR] *and* [GR] *procedures, their discrete modifications and the* [BH] *procedure for* $m = 800$. *Panel (a)–(c) show results for* $m_3 = 80$, $m_1 = 144$ *with* $q_3 = 0.15, 0.25, 0.4$, *panel (d) shows results for* $m_3 = 640$, $m_1 = 80$ *and* $q = 0.4$.

- There is no relevant difference in power between the procedures [DPB] and [DGR].

In addition, Table 4 reports the median FDP of all the procedures. They are all smaller than the nominal level $\alpha = 5\%$, which is in accordance with the theoretical findings. Actually, all procedures are far from exhausting the nominal level. This is due to the strong discreteness induced by the present simulation setting, which makes the testing problem especially difficult. Still, the numerical results of this section show that, by taking into account discreteness, significant improvements are possible while controlling the FDX error rate.

5.3. Analysis of pharmacovigilance data

We revisit the analysis of pharmacovigilance data from Heller and Gur (2011) presented in Döhler et al. (2018). This data set is obtained from a database for reporting, investigating and monitoring adverse drug reactions due to the Medicines and Healthcare products Regulatory Agency in the United Kingdom. It contains the number of reported cases of amnesia as well as the total number of adverse events reported for each of the $m = 2446$ drugs in the database. For a more detailed description of the data which is contained in the R-packages Heller et al. (2012) and Durand and Junge (2019) we refer to Heller and Gur (2011). Heller and Gur (2011) and Döhler et al. (2018) investigate the association between reports of amnesia and suspected drugs by performing for each drug

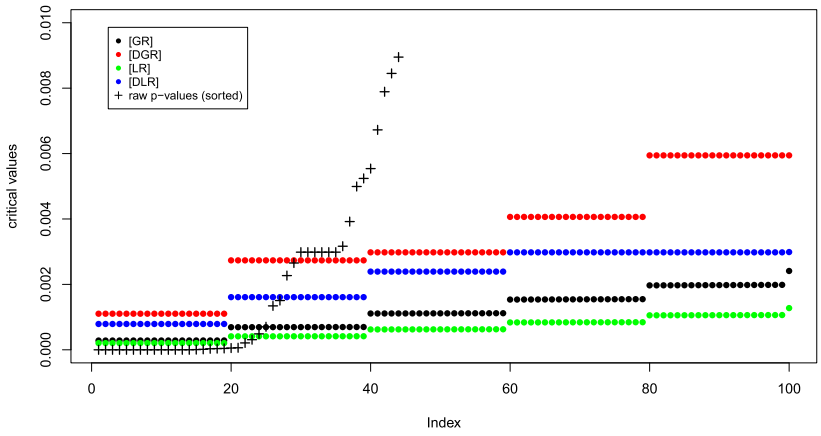

FIG 4. *Critical values and sorted p-values (represented by black crosses) for median FDX control using the* [LR], [DLR], [GR] *and* [DGR] *procedures for the pharmacovigilance data.*

a Fisher's exact test (one-sided) for testing association between the drug and amnesia while adjusting for multiplicity by using several (discrete) FDR controlling procedures. Applying the Benjamini-Hochberg procedure to this data set yields 24 candidate drugs which could be associated with amnesia. Using the discrete FDR controlling procedures from Döhler et al. (2018) yields 27 candidate drugs.

In what follows, we investigate the performance of the [LR], [DLR], [GR], [DPB] and [DGR] procedures for analyzing this data set. First, we compare these procedures when the goal is control of the FDP median instead of FDR at the 5% level, i.e., we require $\mathbf{P}(\text{FDP} > 5\%) \leq 0.5$. Figure 4 illustrates the data and the critical constants of the involved FDX controlling procedures.

The benefit of taking discreteness into account is evident: the discrete critical values are considerably (by a factor of $2.5 \sim 4$) larger than their respective classical counterparts which leads to more powerful procedures, see also Table 2.

TABLE 2
Number of rejections for the pharmacovigilance data.

Procedure controls	[LR]	[DLR]	[GR]	[DPB]	[DGR]
$\mathbf{P}(\text{FDP} > 5\%) \leq 0.5$	23	27	24	29	29
$\mathbf{P}(\text{FDP} > 5\%) \leq 0.05$	16	21	16	24	24

Note that the critical values of [DPB] are not displayed in Figure 4 since they are visually indistinguishable from the [DGR] critical values. Figure 5 shows that this is in fact true for all indices, thus [DGR] is not only an efficient, but also quite accurate approximation of the [DPB] values, at least for the discrete distribution involved in this example.

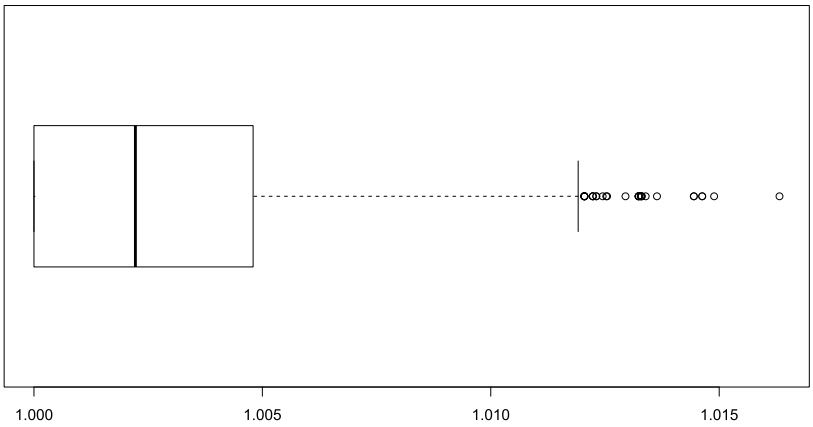

FIG 5. *Boxplot for the ratio of the* [DPB] *to the* [DGR] *critical values.*

We also compare the performance of the above procedures over the full range of possible values for ζ. Figure 6 depicts the number of rejections when controlling $\mathbf{P}(\text{FDP} > 5\%) \leq \zeta$ for $\zeta \in (0, 1)$. As expected from Propositions 3.1, 3.2 and 3.3, the discrete variants reject more hypotheses than their classical counterparts for all values of ζ. For central values of ζ, the gain is about three to four additionally rejected hypotheses, which corresponds roughly to the gain from using the discrete version of [BH] instead of [BH] (see Table 1 in Döhler et al. (2018)). Figure 6 also shows that for more extreme values of ζ the gain may be more pronounced, e.g., when $\mathbf{P}(\text{FDP} > 5\%) \leq 0.05$ is to be guaranteed, the [GR] procedure rejects 16 hypotheses, whereas the [DGR] procedure rejects 24 hypotheses (see the second row of Table 2).

6. Discussion

In this paper, we presented new procedures controlling the FDX while incorporating the (heterogeneous) family of null distribution $\{F_i, 1 \leq i \leq m\}$. Markedly, it put forward that the geometric averaging of the F_i's is a suitable operation for FDX control. This is new to our knowledge, as all previous works are mostly based on arithmetic averaging of the F_i's (or variation thereof). Maybe more importantly, our approach led to a substantial power improvement in two common situations, under continuity of the test statistics via weighting schemes, and for discrete test statistics when performing multiple individual Fisher's exact tests.

This work opens several directions of research. First, the proofs of all our FDX bounds rely on using a kind of independence between the p-values (see (Indep0) and (Indep)). While this assumption is classical, it is desirable to remove this condition in order to obtain more realistic models. This generalization seems however challenging, as FDX control under dependence is already delicate to

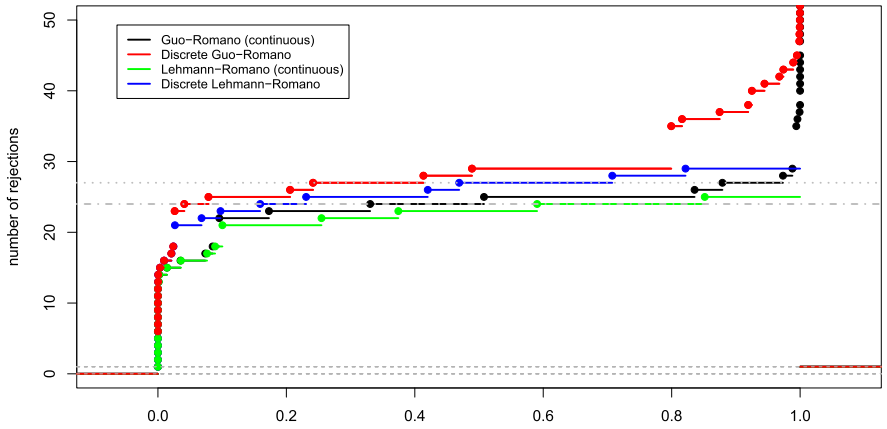

FIG 6. *Number of rejections (y-axis) for the pharmacovigilance data when using the* [LR], [DLR], [GR] *and* [DGR] *procedures. All procedures control the tail probabilities at level* ζ *(on the x-axis) for the* FDP *exceeding* 5%. *The horizontal lines represent the rejections of the BH and discrete BH procedures at* FDR-*level* 5%.

study in the homogeneous case, see Delattre and Roquain (2015). A second interesting avenue is to derive theoretical bounds for the true discovery proportion (TDP) of our procedure. In particular, a relevant concern is to assess whether our way of accounting for heterogeneity (via arithmetic or geometric averaging of the F_i's) is optimal in some sense. Lastly, our work paves the way to control other simultaneous inference criteria based on an event probability, e.g., to establish post hoc bounds in the discrete heterogeneous case, see Genovese and Wasserman (2006); Goeman and Solari (2011); Blanchard et al. (2020). While challenging, this is a very exciting direction for future research.

Finally, we mention that the FDX procedures described in Sections 4 and 5 have been implemented in the R package `FDX`, see Junge and Döhler (2020).

Appendix A: Materials for the proofs

A.1. Proving the main tool

The proof is based on the following result, which is a reformulation of Theorem 5.2 in Roquain (2011) in our context.

Theorem A.1 (Roquain, 2011)**.** *In the setting defined in Section 2.1, consider any step-down procedure* R *with critical values* τ_ℓ, $1 \leq \ell \leq m$. *Then for all* $P \in \mathcal{P}$, *we have*

$$\mathrm{FDX}(R,P) \le \sum_{\ell=1}^{m} \mathbb{1}\{|\mathcal{H}_0(P)| \le m(\ell)\} \times \mathbf{P}_{X\sim P}\left(\sum_{i\in\mathcal{H}_0(P)} \mathbb{1}\{p_i(X)\le \tau_\ell\} \ge \lfloor \alpha\ell \rfloor + 1, \widetilde{\ell}(P) = \ell \right), \quad (34)$$

where $\widetilde{\ell}(P) = \min\left\{ \ell \in \{1,\dots,m\} \,:\, \ell - \sum_{i\in\mathcal{H}_1(P)} \mathbb{1}\{p_i(X)\le\tau_\ell\} \ge \lfloor\alpha\ell\rfloor + 1 \right\}$ *(with* $\widetilde{\ell}(P) = m+1$ *if the set is empty).*

Let us show that Theorem A.1 implies (17) under (Indep0). Under (Indep0), $\widetilde{\ell}(P)$ is independent of the random variable family

$$\left\{ \sum_{i\in\mathcal{H}_0(P)} \mathbb{1}\{p_i(X)\le\tau_\ell\},\ 1\le\ell\le m \right\}.$$

Hence, (34) provides that $\mathrm{FDX}(R,P)$ is smaller or equal to

$$\sum_{\ell=1}^{m} \mathbb{1}\{|\mathcal{H}_0(P)|\le m(\ell)\}\ \mathbf{P}_{X\sim P}\left(\sum_{i\in\mathcal{H}_0(P)} \mathbb{1}\{p_i(X)\le\tau_\ell\} \ge \lfloor\alpha\ell\rfloor+1 \right) \times \mathbf{P}_{X\sim P}\left(\widetilde{\ell}(P)=\ell\right) \le B(\tau,\alpha),$$

which gives (17).

Finally, for completeness, let us now prove Theorem A.1. Let $R_\ell = \{i \in \{1,\dots,m\} \,:\, p_i(X)\le\tau_\ell\}$ for all ℓ. First, we have for any $\ell\in\{1,\dots,m\}$ such that $|R_\ell| = \ell$:

$$\begin{aligned}\{\mathrm{FDP}(R_\ell,P) > \alpha\} &= \{|\mathcal{H}_0(P)\cap R_\ell| > \alpha\ell\} = \{|\mathcal{H}_0(P)\cap R_\ell| \ge \lfloor\alpha\ell\rfloor+1\} \\ &= \{\ell - |\mathcal{H}_1(P)\cap R_\ell| \ge \lfloor\alpha\ell\rfloor+1\} \subset \{\ell \ge \widetilde{\ell}(P)\},\end{aligned}$$

by using the definition of $\widetilde{\ell}(P)$. Assuming now $|R_{\ell'}|\ge\ell'$ for all $\ell'\le\ell$, we obtain

$$\begin{aligned}\{\mathrm{FDP}(R_\ell,P)>\alpha\} &\subset \{\ell\ge\widetilde{\ell}(P), |R_{\widetilde{\ell}(P)}|\ge\widetilde{\ell}(P)\} \\ &\subset \{|\mathcal{H}_0(P)\cap R_{\widetilde{\ell}(P)}|\ge\lfloor\alpha\widetilde{\ell}(P)\rfloor+1\},\end{aligned}$$

where the last step uses the definition of $\widetilde{\ell}(P)$. Moreover, if $\widetilde{\ell}(P)\ge 2$, by minimality of $\widetilde{\ell}(P)$, we have $(\widetilde{\ell}(P)-1) - \sum_{i\in\mathcal{H}_1(P)} \mathbb{1}\{p_i(X)\le\tau_{\widetilde{\ell}(P)-1}\} < \lfloor\alpha(\widetilde{\ell}(P)-1)\rfloor+1$. Hence, we obtain the following upper-bound for $|\mathcal{H}_0(P)|$:

$$|\mathcal{H}_0(P)| = m - |\mathcal{H}_1(P)| \le m - |\mathcal{H}_1(P)\cap R_{\widetilde{\ell}(P)-1}| \le m - \widetilde{\ell}(P) + \lfloor\alpha(\widetilde{\ell}(P)-1)\rfloor + 1.$$

Since the above bound is also true when $\widetilde{\ell}(P)=1$, it holds for any possible value of $\widetilde{\ell}(P)\le m$. Since $\ell = \widehat{\ell}$ in (9) satisfies both $|R_\ell|=\ell$ and $|R_{\ell'}|\ge\ell'$ for all $\ell'\le\ell$, combining the above displays gives (34).

A.2. Proof of Proposition 2.1

First, we have with P-probability 1, for all $i \in \{1,\dots,m\}$, $p_i \in \mathcal{A}$, both under (Cont) or (Discrete). Hence, by (10), we have $\{\ell \in \{1,\dots,m\} : \xi_\ell(p_{\sigma(\ell)}) \le \zeta\} = \{\ell \in \{1,\dots,m\} : p_{\sigma(\ell)} \le \tau_\ell\}$. By (9), this gives

$$\widehat{\ell} = \max\{\ell \in \{0,\dots,m\} : \forall \ell' \le \ell,\ p'_{\ell'} \le \zeta\},$$

where we have denoted $p'_\ell = \xi_\ell(p_{\sigma(\ell)})$ for all ℓ. Now note that

$$\begin{aligned}\{\sigma(1),\dots,\sigma(\widehat{\ell})\} &= \{i \in \{1,\dots,m\} : \sigma^{-1}(i) \in \{1,\dots,\widehat{\ell}\}\}\\ &= \{i \in \{1,\dots,m\} : \forall \ell \in \{1,\dots,\sigma^{-1}(i)\}, p'_\ell \le \zeta\}\\ &= \{i \in \{1,\dots,m\} : \max_{\ell \in \{1,\dots,\sigma^{-1}(i)\}} \{p'_\ell\} \le \zeta\},\end{aligned}$$

hence it is sufficient to prove that $\tilde{p}_i = \max_{\ell \in \{1,\dots,\sigma^{-1}(i)\}}\{p'_\ell\}$ for any $i \in \{1,\dots,m\}$. For this, let us fix $i \in \{1,\dots,m\}$ and write $\{\ell \in \{1,\dots,m\} : p_{\sigma(\ell)} \le p_i\} = \{\ell \in \{1,\dots,m\} : \ell \le \sigma^{-1}(i)\} \cup A$, for $A = \{\ell \in \{1,\dots,m\} : p_{\sigma(\ell)} \le p_i, \ell > \sigma^{-1}(i)\}$. This is possible because, by definition, $\ell \le \sigma^{-1}(i)$ implies $p_{\sigma(\ell)} \le p_i$. Next, for any $\ell \in A$, we have both $p_{\sigma(\ell)} \le p_i$ and $p_{\sigma(\ell)} \ge p_i$, which entails $p_{\sigma(\ell)} = p_i$ and thus $\xi_\ell(p_{\sigma(\ell)}) = \xi_\ell(p_i)$. Since $\sigma^{-1}(i) \le \ell$ and by the nonincreasing property of $\ell \mapsto \xi_\ell(p_i)$, we have $\xi_\ell(p_i) \le \xi_{\sigma^{-1}(i)}(p_i) = \xi_{\sigma^{-1}(i)}(p_{\sigma(\sigma^{-1}(i))})$. This gives $p'_\ell \le p'_{\sigma^{-1}(i)}$ for all $\ell \in A$. Therefore,

$$\max_{\substack{1\le\ell\le m\\ p_{\sigma(\ell)}\le p_i}} \{p'_\ell\} = \max_{\substack{1\le\ell\le m\\ \ell\le\sigma^{-1}(i)}} \{p'_\ell\} \vee \max_{\ell\in A}\{p'_\ell\} = \max_{\substack{1\le\ell\le m\\ \ell\le\sigma^{-1}(i)}} \{p'_\ell\},$$

which leads to the result.

A.3. Auxiliary lemmas

Lemma A.1. *With the notation in (26) the quantity*

$$\mathbf{P}\left(\mathbf{Bin}\left[m-\ell+i, \tilde{F}_{m-\ell+i}(t)\right] \ge i\right), \tag{35}$$

which is equal to $\mathbf{P}\left(\mathbf{Bin}\left[m-\ell+i, 1-\tilde{F}_{m-\ell+i}(t)\right] \le m-\ell\right)$*, is non-increasing both in* $i \in \{1,\dots,\lfloor \alpha\ell \rfloor + 1\}$ *and* $\ell \in \{1,\dots,m\}$.

Proof. First note that

$$1-\tilde{F}_j(t) = \left(\prod_{j'=1}^{j}(1-F_{(j')}(t))\right)^{1/j}$$

in non-decreasing in j (because the geometric average of larger numbers is larger). The quantity (35) is thus non-increasing with respect to i, so that the

only thing to check is that this quantity is non-increasing with respect to ℓ. For this, it is sufficient to prove that $\mathbf{Bin}\left[j+1, \tilde{F}_{j+1}(t)\right]$ is stochastically larger than $\mathbf{Bin}\left[j, \tilde{F}_j(t)\right]$ for any $j \in \{1, \dots, m-1\}$ (which is not obvious because $\tilde{F}_j(t) \geq \tilde{F}_{j+1}(t)$). Let $n_1 = j$, $p_1 = \tilde{F}_j(t)$, $n_2 = 1$, $p_2 = F_{(j+1)}(t)$, $n = j+1$ and $p = \tilde{F}_{j+1}(t)$. We easily check that $n = n_1 + n_2$ and by (26),

$$\begin{aligned}(1-p)^n &= \prod_{j'=1}^{j+1}(1-F_{(j')}(t)) \\ &= \prod_{j'=1}^{j}(1-F_{(j')}(t)) \times (1-F_{(j+1)}(t)) = (1-p_1)^{n_1}(1-p_2)^{n_2}.\end{aligned}$$

Applying Example 1.A.25 in Shaked, M. and Shanthikumar, J.G. (2007) ($m = 2$ with the notation therein), we obtain that the sum of a $\mathbf{Bin}[n_1, p_1]$ variable and a $\mathbf{Bin}[n_2, p_2]$ variable (with independence) is stochastically smaller than a $\mathbf{Bin}[n, p]$ variable. In particular, a $\mathbf{Bin}[n_1, p_1]$ variable is stochastically smaller than a $\mathbf{Bin}[n, p]$ variable. This gives the result. □

Lemma A.2. *For any $x, x' \in \mathbb{R}^m$, we have*

$$\sup_{1\leq k\leq m} |x_{(k)} - x'_{(k)}| \leq \sup_{1\leq i\leq m} |x_i - x'_i|, \tag{36}$$

where $x_{(1)} \geq \cdots \geq x_{(m)}$ and $x'_{(1)} \geq \cdots \geq x'_{(m)}$.

Proof. Let σ (resp. σ') be a permutation of $\{1, \dots, m\}$ ordering x (resp. x'), that is, such that $x_{(k)} = x_{\sigma(k)}$ (resp. $x'_{(k)} = x'_{\sigma'(k)}$) for all $k \in \{1, \dots, m\}$. Let $k \in \{1, \dots, m\}$ and prove $|x_{\sigma(k)} - x'_{\sigma'(k)}| \leq \sup_{1\leq i\leq m} |x_i - x'_i|$. Let us assume without loss of generality that $x_{\sigma(k)} \geq x'_{\sigma'(k)}$. On the one hand, if $\sigma'(k) \in \{\sigma(1), \dots, \sigma(k)\}$, then $x_{\sigma(k)} - x'_{\sigma'(k)} \leq x_{\sigma'(k)} - x'_{\sigma'(k)} \leq \sup_{1\leq i\leq m} |x_i - x'_i|$. On the other hand, if $\sigma'(k) \notin \{\sigma(1), \dots, \sigma(k)\}$, then $\sigma'(k) \in \{\sigma(k+1), \dots, \sigma(m)\}$ and $\sigma'(k+1), \dots, \sigma'(m)$ cannot all fall into $\{\sigma(k+1), \dots, \sigma(m)\}$. This shows that there exists a $j_0 > k$ with $\sigma'(j_0) \in \{\sigma(1), \dots, \sigma(k)\}$. Hence, $x_{\sigma(k)} - x'_{\sigma'(k)} \leq x_{\sigma'(j_0)} - x'_{\sigma'(k)} \leq x_{\sigma'(j_0)} - x'_{\sigma'(j_0)}$ (because $k \leq j_0$ entails $x'_{\sigma'(k)} \geq x'_{\sigma'(j_0)}$), which can be further upper-bounded by $\sup_{1\leq i\leq m} |x_i - x'_i|$. □

Appendix B: Additional numerical details

TABLE 3
Average power (i.e. average of true discovery proportion) of median FDP controlling procedures ($\zeta = 0.5$) for $\alpha = 5\%$ and $N = 25$.

m	m_3	m_1	q_3	[BH]	[LR]	[DLR]	[GR]	[DPB]	[DGR]
800	80	144	0.15	0.0000	0.0001	0.0025	0.0002	0.0025	0.0025
		144	0.25	0.0004	0.0031	0.0430	0.0077	0.0430	0.0430
		144	0.40	0.0803	0.0853	0.3328	0.1195	0.4412	0.4406
		360	0.15	0.0000	0.0001	0.0025	0.0002	0.0043	0.0043
		360	0.25	0.0004	0.0031	0.0430	0.0077	0.0444	0.0444
		360	0.40	0.0803	0.0853	0.3766	0.1195	0.4512	0.4511
		576	0.15	0.0000	0.0001	0.0071	0.0002	0.0076	0.0076
		576	0.25	0.0004	0.0031	0.0528	0.0077	0.0770	0.0770
		576	0.40	0.0803	0.0853	0.4474	0.1195	0.5141	0.5128
	240	112	0.15	0.0000	0.0000	0.0025	0.0002	0.0025	0.0025
		112	0.25	0.0005	0.0031	0.0289	0.0076	0.0422	0.0422
		112	0.40	0.2148	0.1226	0.4250	0.1984	0.5153	0.5139
		280	0.15	0.0000	0.0000	0.0025	0.0002	0.0025	0.0025
		280	0.25	0.0005	0.0031	0.0336	0.0076	0.0429	0.0429
		280	0.40	0.2147	0.1226	0.4413	0.1983	0.5728	0.5716
		448	0.15	0.0000	0.0000	0.0025	0.0002	0.0037	0.0037
		448	0.25	0.0005	0.0031	0.0389	0.0076	0.0430	0.0430
		448	0.40	0.2145	0.1226	0.4609	0.1983	0.5921	0.5917
	640	32	0.15	0.0000	0.0000	0.0018	0.0002	0.0025	0.0025
		32	0.25	0.0010	0.0031	0.0203	0.0075	0.0212	0.0212
		32	0.40	0.4243	0.3009	0.4908	0.5379	0.6730	0.6724
		80	0.15	0.0000	0.0000	0.0020	0.0002	0.0025	0.0025
		80	0.25	0.0010	0.0031	0.0203	0.0075	0.0212	0.0212
		80	0.40	0.4242	0.3008	0.4974	0.5374	0.6746	0.6743
		128	0.15	0.0000	0.0000	0.0021	0.0002	0.0025	0.0025
		128	0.25	0.0010	0.0031	0.0203	0.0075	0.0212	0.0212
		128	0.40	0.4240	0.3008	0.5048	0.5369	0.6753	0.6750
2000	200	360	0.15	0.0000	0.0000	0.0007	0.0000	0.0022	0.0022
		360	0.25	0.0001	0.0024	0.0198	0.0029	0.0222	0.0222
		360	0.40	0.0730	0.0560	0.3331	0.0792	0.4315	0.4311
		900	0.15	0.0000	0.0000	0.0022	0.0000	0.0024	0.0024
		900	0.25	0.0001	0.0024	0.0210	0.0029	0.0373	0.0373
		900	0.40	0.0730	0.0560	0.3380	0.0792	0.4515	0.4515
		1440	0.15	0.0000	0.0000	0.0024	0.0000	0.0024	0.0024
		1440	0.25	0.0001	0.0024	0.0378	0.0029	0.0428	0.0428
		1440	0.40	0.0729	0.0560	0.4320	0.0792	0.5173	0.5144
	600	280	0.15	0.0000	0.0000	0.0007	0.0000	0.0007	0.0007
		280	0.25	0.0001	0.0024	0.0197	0.0029	0.0205	0.0205
		280	0.40	0.2058	0.1137	0.4093	0.1960	0.5194	0.5176
		700	0.15	0.0000	0.0000	0.0007	0.0000	0.0020	0.0020
		700	0.25	0.0001	0.0024	0.0200	0.0029	0.0205	0.0205
		700	0.40	0.2058	0.1137	0.4374	0.1960	0.5678	0.5657
		1120	0.15	0.0000	0.0000	0.0014	0.0000	0.0024	0.0024
		1120	0.25	0.0001	0.0024	0.0201	0.0029	0.0206	0.0206
		1120	0.40	0.2057	0.1137	0.4545	0.1959	0.5908	0.5906
	1600	80	0.15	0.0000	0.0000	0.0007	0.0000	0.0007	0.0007
		80	0.25	0.0003	0.0024	0.0090	0.0029	0.0172	0.0172
		80	0.40	0.4223	0.2949	0.4823	0.5288	0.6665	0.6658
		200	0.15	0.0000	0.0000	0.0007	0.0000	0.0007	0.0007
		200	0.25	0.0003	0.0024	0.0090	0.0029	0.0184	0.0184
		200	0.40	0.4222	0.2949	0.4866	0.5286	0.6689	0.6683
		320	0.15	0.0000	0.0000	0.0007	0.0000	0.0007	0.0007
		320	0.25	0.0003	0.0024	0.0090	0.0029	0.0194	0.0194
		320	0.40	0.4220	0.2948	0.4935	0.5283	0.6724	0.6715

TABLE 4
Estimated median FDP values for $\alpha = 5\%$ and $N = 25$.

m	m_3	m_1	q_3	[BH]	[LR]	[DLR]	[GR]	[DPB]	[DGR]
800	80	144	0,15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		144	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		144	0.40	0.0000	0.0000	0.0000	0.0000	0.0014	0.0014
		360	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		360	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		360	0.40	0.0000	0.0000	0.0000	0.0000	0.0014	0.0014
		576	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		576	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		576	0.40	0.0000	0.0000	0.0000	0.0000	0.0014	0.0014
	240	112	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		112	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		112	0.40	0.0000	0.0000	0.0018	0.0000	0.0036	0.0036
		280	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		280	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		280	0.40	0.0000	0.0000	0.0000	0.0000	0.0036	0.0036
		448	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		448	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		448	0.40	0.0000	0.0000	0.0000	0.0000	0.0018	0.0018
	640	32	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		32	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		32	0.40	0.0000	0.0000	0.0062	0.0062	0.0125	0.0125
		80	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		80	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		80	0.40	0.0000	0.0000	0.0000	0.0000	0.0062	0.0062
		128	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		128	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		128	0.40	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2000	200	360	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		360	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		360	0.40	0.0000	0.0000	0.0000	0.0000	0.0011	0.0011
		900	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		900	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		900	0.40	0.0000	0.0000	0.0000	0.0000	0.0006	0.0006
		1440	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		1440	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		1440	0.40	0.0000	0.0000	0.0000	0.0000	0.0011	0.0011
	600	280	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		280	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		280	0.40	0.0000	0.0000	0.0014	0.0000	0.0043	0.0043
		700	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		700	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		700	0.40	0.0000	0.0000	0.0007	0.0000	0.0029	0.0029
		1120	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		1120	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		1120	0.40	0.0000	0.0000	0.0007	0.0000	0.0021	0.0021
	1600	80	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		80	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		80	0.40	0.0000	0.0000	0.0050	0.0050	0.0150	0.0150
		200	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		200	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		200	0.40	0.0000	0.0000	0.0025	0.0025	0.0100	0.0100
		320	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		320	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		320	0.40	0.0000	0.0000	0.0000	0.0000	0.0025	0.0025

Acknowledgements

This work has been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS) and by the GDR ISIS through the "projets exploratoires" program (project TASTY). It is part of project DO 2463/1-1, funded by the Deutsche Forschungsgemeinschaft. The authors thank Florian Junge for implementing the heterogeneous FDX procedures and improved Poisson-binomial distribution functions in R, and for running the simulations.

References

Basu, P., Fu, L., Saretto, A., and Sun, W. (2020). Empirical bayes control of the false discovery exceedance. Personal communication.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300. MR1325392

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, 48(3):1281–1303. MR4124323

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992. MR2448601

Chen, X., Doerge, R. W., and Heyse, J. F. (2018). Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biom. J.*, 60(4):761–779. MR3830957

Chen, X., Doerge, R. W., and Sarkar, S. K. (2020). A weighted FDR procedure under discrete and heterogeneous null distributions. *Biom. J.*, 62(6):1544–1563. MR4164153

Delattre, S. and Roquain, E. (2011). On the false discovery proportion convergence under Gaussian equi-correlation. *Statist. Probab. Lett.*, 81(1):111–115. MR2740072

Delattre, S. and Roquain, E. (2015). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *Ann. Statist.*, 43(3):1141–1177. MR3346700

Delattre, S. and Roquain, E. (2016). On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing. *Bernoulli*, 22(1):302–324. MR3449784

Dickhaus, T., Straßburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T., and Navarro, A. (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical applications in genetics and molecular biology*, 11(4). MR2958611

Ditzhaus, M. and Janssen, A. (2019). Variability and stability of the false discovery proportion. *Electron. J. Statist.*, 13(1):882–910. MR3932610

Döhler, S. (2016). A discrete modification of the Benjamini—Yekutieli procedure. *Econometrics and Statistics*. MR3740116

Döhler, S., Durand, G., and Roquain, E. (2018). New FDR bounds for discrete and heterogeneous tests. *Electron. J. Statist.*, 12(1):1867–1900. MR3813600

Dudoit, S. and van der Laan, M. J. (2007). *Multiple Testing Procedures and Applications to Genomics*. Springer Series in Statistics. Springer. ISBN: 978-0-387-49316-9. MR2373771

Durand, G. (2019). Adaptive p-value weighting with power optimality. *Electron. J. Statist.*, 13(2):3336–3385. MR4010982

Durand, G. and Junge, F. (2019). *DiscreteFDR: Multiple Testing Procedures with Adaptation for Discrete Tests*. R package version 1.2.

Durand, G., Junge, F., Döhler, S., and Roquain, E. (2019). DiscreteFDR: An R package for controlling the false discovery rate for discrete test statistics. *arXiv e-prints*, arXiv:1904.02054. MR4122305

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061. MR2065197

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524. MR2261439

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417. MR2279468

Gilbert, P. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society. Series C*, 54(1):143–158. MR2134603

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597. MR2951390

Guo, W., He, L., and Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.*, 42(3):1070–1101. MR3210996

Guo, W. and Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 3, 35 pp. (electronic). MR2306938

Habiger, J. D. (2015). Multiple test functions and adjusted p-values for test statistics with discrete distributions. *J. Statist. Plann. Inference*, 167:1–13. MR3383232

Heller, R. and Gur, H. (2011). False discovery rate controlling procedures for discrete tests. *ArXiv e-prints*.

Heller, R., Gur, H., and Yaacoby, S. (2012). *discreteMTP: Multiple testing procedures for discrete test statistics*. R package version 0.1-2.

Hemerik, J., Solari, A., and Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649. MR3992394

Heyse, J. F. (2011). A false discovery rate procedure for categorical data. In *Recent Advances in biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70. MR0538597

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.*, 105(491):1215–1227. MR2752616

Ignatiadis, N., Klaus, B., Zaugg, J., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13:577–580.

Junge, F. (2020). *PoissonBinomial: Efficient Computation of Ordinary and Generalized Poisson Binomial Distributions*. R package version 1.1.1.

Junge, F. and Döhler, S. (2020). *FDX: FDX Controlling Multiple Testing Procedures for Heterogeneous and Discrete Tests*. R package version 0.1-3.

Katsevich, E. and Ramdas, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression, and online settings. *Ann. Statist.* to appear.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Statist. Plann. Inference*, 124(2):379–398. MR2080371

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.*, 33:1138–1154. MR2195631

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Stat.*, 2:1065–1110. MR2460858

Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.*, 99(468):1002–1014. MR2109490

Ramdas, A., Foygel Barber, R., Wainwright, M. J., and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.*, 47(5):2790–2821. MR3988773

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408. MR2351090

Roquain, E. (2011). Type I error rate control for testing many hypotheses: a survey with proofs. *J. Soc. Fr. Stat.*, 152(2):3–38. MR2821220

Roquain, E. and van de Wiel, M. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.*, 3:678–711. MR2521216

Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.*, 39(1):584–612. MR2797857

Rubin, D., Dudoit, S., and van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 19, 20 pp. (electronic). MR2240850

Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*. Springer Series in Statistics. Springer New York. MR2265633

Tan, X., Liu, G. F., Zeng, D., Wang, W., Diao, G., Heyse, J. F., and Ibrahim, J. G. (2019). Controlling false discovery proportion in identification of drug-related adverse events from multiple system organ classes. *Statistics in Medicine*, 38(22):4378–4389. MR4002437

Tarone, R. E. (1990). A modified bonferroni method for discrete data. *Biometrics*, 46(2):515–522.

Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University.

Westfall, P. and Wolfinger, R. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51(1):3–8.

Zhao, H. and Zhang, J. (2014). Weighted p-value procedures for controlling FDR of grouped hypotheses. *J. Statist. Plann. Inference*, 151/152:90–106. MR3216640