Received: 24 January 2022 | Revised: 24 November 2022 | Accepted: 18 March 2023

DOI: 10.1002/sim.9725

RESEARCH ARTICLE

Statistics in Medicine WILEY

Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis

Angela Andreella[1] | **Jesse Hemerik**[2] | **Livio Finos**[3] | **Wouter Weeda**[4] | **Jelle Goeman**[5]

[1]Department of Economics, Ca' Foscari University of Venice, Venice, Italy

[2]Biometris, Wageningen University and Research, Wageningen, The Netherlands

[3]Department of Statistics, University of Padova, Padova, Italy

[4]Department of Psychology, Leiden University, Leiden, The Netherlands

[5]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Angela Andreella, Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy.
Email: angela.andreella@unive.it

Summary

We propose a permutation-based method for testing a large collection of hypotheses simultaneously. Our method provides lower bounds for the number of true discoveries in any selected subset of hypotheses. These bounds are simultaneously valid with high confidence. The methodology is particularly useful in functional Magnetic Resonance Imaging cluster analysis, where it provides a confidence statement on the percentage of truly activated voxels within clusters of voxels, avoiding the well-known spatial specificity paradox. We offer a user-friendly tool to estimate the percentage of true discoveries for each cluster while controlling the family-wise error rate for multiple testing and taking into account that the cluster was chosen in a data-driven way. The method adapts to the spatial correlation structure that characterizes functional Magnetic Resonance Imaging data, gaining power over parametric approaches.

KEYWORDS

fMRI cluster analysis, multiple testing, permutation test, selective inference, true discovery proportion

1 | INTRODUCTION

Functional Magnetic Resonance Imaging (fMRI) is the most frequently used technique to understand which regions of the human brain are activated as a consequence of a stimulus. Brain activation is measured as the correlation between a sequence of (cognitive) stimuli and the resulting blood oxygenation level dependent (BOLD) signal. The BOLD signal over the entire brain is measured in small cubes termed voxels, and for each of these, we test for significant BOLD activity. Typically, around 300,000 voxels are analyzed, so the resulting multiple testing problem has roughly 300,000 statistical tests.

Controlling Type I error at the voxel level usually negatively affects the power to detect activation.[1] Therefore, cluster-extent based thresholding was developed to analyze the data at the level of clusters of contiguous voxels. This method is less conservative than voxel-wise inference since it exploits the spatial nature of the signal using Random Field Theory (RFT).[2] However, the assumptions behind this method require a very high initial cluster-forming threshold, resulting in relatively small clusters left for significance testing.[3] Moreover, the method suffers from the spatial specificity

Abbreviations: AORC, asymptotically optimal rejection curves; ARI, all-resolution inference; BET, brain extraction tool; BOLD, blood oxygenation level dependent; FDR, false discovery rate; FLIRT, FMRIB's linear image registration tool; fMRI, functional Magnetic Resonance; FSL, FMRIB software library; FWER, family-wise error rate; FWHM, full width at half maximum; MNI, Montreal neurological institute; MCFLIRT, motion correction FLIRT; PRDS, positive regression dependency on subset; RFT, random field theory; SPM, statistical parametric mapping; TDP, true discovery proportion; TFCE, threshold-free cluster enhancement.

paradox,[4] meaning that the larger the cluster we find, the less we can say about the signal within that cluster. Since the method tests the hypothesis that none of the voxels in the cluster are active, rejecting this null hypothesis only allows the claim that there is at least one active voxel inside the cluster; hence, the larger the cluster, the less we can say about it. That is, the number of active voxels and their spatial location remains unknown, and doing follow-up inference inside the cluster ("drilling down") leads to a "double-dipping" problem and inflated Type I error rate.[5]

These problems motivated Rosenblatt et al[6] to propose All-Resolution Inference (ARI), a method to compute the lower confidence bound for the true number of active voxels within a cluster (true discovery proportion (TDP)) simultaneously for all possible sets, for example, all clusters of voxels. Simultaneous control permits users to drill down within clusters while maintaining error guarantees, thus resolving the spatial specificity paradox. ARI is based on the approach proposed by Goeman and Solari[7] using closed testing[8] with local Simes test[9] to control the family-wise error rate (FWER). The closed testing method has an exponential computational load in general; nevertheless, for this specific case, Goeman et al[10] and Meijer et al[11] proposed a fast and exact linear time short-cut. ARI relies on the Simes inequality, assuming positive regression dependency on subsets (PRDS).[12] While the Simes inequality can be assumed to be valid for fMRI data,[1] it can be conservative under strong positive dependence. This makes the method inefficient in the neuroimaging data framework since brain measurements have strong spatial dependence due to both physics and physiology.

Permutation tests assume only exchangeability under the null hypothesis[13] and can handle data having any correlation structure, adapting to that correlation structure both to keep type I error control and to gain power. Hemerik et al[14] proposed a permutation-based method, related to ARI, that adapts the procedure to the correlation structure of the p-values. However, this method finds TDP only for sets consisting of the smallest k p-values, simultaneously over k, and can therefore not handle spatially defined clusters. Moreover, the method allows much freedom in the choice of the shape of its rejection curve, and the optimal choice for fMRI data is not clear.

In this paper, we merge the strengths of ARI with the permutation-based method of Hemerik et al,[14] adapting ARI to use permutations following the approach of Hemerik et al.[14] The new method provides a lower bound for the TDP for all brain regions, allowing regions of interest to be chosen post-hoc, as in ARI, without compromising family-wise error control. By using permutation-based test statistics, the method gains in power compared to the parametric version of ARI because it adapts to the correlation structure. Moreover, permutation tests are robust, as widely demonstrated in the neuroimaging literature[3,15,16] and can be used when the parametric assumptions of ARI are not satisfied. The permutation-based post-hoc method proposed here is similar to the one offered by Blanchard et al,[17] which essentially generalized the approach of Hemerik et al[14] to arbitrary subsets of the hypotheses. However, we also propose here an iterative approach based on the idea presented by Hemerik et al,[14] which uniformly improves Blanchard et al[17] method in most cases.

This paper is organized as follows. Section 2 introduces the concept of closed testing based on a critical vector, revisiting the results from Goeman et al[18] and Rosenblatt et al.[6] Then, in Section 3, we combine these results to obtain a permutation-based ARI, and its iterative version in Section 4. We discuss the families of critical curves to be used in Section 5 and which test and permutations we recommend to use in fMRI data in Section 6. Section 7 evaluates the performance of our method in comparison with the parametric version in fMRI data. We validate the method using the resting-state fMRI null data of Eklund et al[3] in Section 8. Finally, we perform some simulations in Section 9 in order to investigate the influence of the shape of the rejection curve in different scenarios.

2 | CLOSED TESTING FOR TRUE DISCOVERY PROPORTIONS

In this section, we revisit some results from Rosenblatt et al[6] and Goeman et al[18] to introduce notation and clarify the need for selective inference in fMRI data.

Suppose the brain B, with $|B| = m$, is composed of m voxels, and let 2^B be the collection of all subsets of the brain. Some of the voxels are truly active: let $A \subseteq B$ be the unknown set of all truly active voxels. For a cluster of interest $S \subseteq B$ we want to make inference on $a(S) = |A \cap S|$, that is, the number of truly active voxels in S, or equivalently the TDP, that is, $a(S)/|S|$.

We assume that we have computed a test statistic for each voxel i, where $i = 1, \dots, m$, corresponding to the null hypothesis that the voxel is not active. Based on some knowledge or guess of the marginal null distribution of these test statistics, we may compute the corresponding (parametric) p-values $p_i : \Omega \to [0, 1]$ where Ω is the sample space of the data X. In the parametric version, we will assume that these p-values will be valid, that is, stochastically smaller than the uniform distribution of all inactive voxels. For the permutation-based method, we emphasize

here that, to guarantee FWER control, we do not make any assumptions on the distribution of these p-values. The reason is that it will suffice that the p-values are computed in the same way for all permuted versions of the data.[14]

We will now revisit the simultaneous inference on TDP using closed testing and critical vectors. First, we define a critical vector.

Definition 1. A vector $(l_1, \dots, l_m)$ is a critical vector if and only if

$$\Pr(\cap_{i=1}^{|N|}\{q_{(i)} \geq l_i\}) \geq 1-\alpha, \tag{1}$$

where $N = B \setminus A$ is the set of inactive voxels, and $q_{(i)}$, $1 \leq i \leq |N|$, are their sorted p-values.

The general parametric version of ARI assumes that, for a chosen error rate $\alpha \in [0,1]$, there is a critical vector $(l_1, \dots, l_m)$, possibly random, expressed as Definition 1. If such a critical vector exists, then as a corollary to Lemma 6 from Goeman et al[18] we have the following theorem, which we prove in Appendix.

Theorem 1. *Let l_i satisfy (1). Then for every $\emptyset \neq S \subseteq B$,*

$$\bar{a}(S) = \max_{1\leq u\leq |S|} 1 - u + |\{i \in S : p_i \leq l_u\}| \tag{2}$$

is a lower $(1-\alpha)$ confidence bound of $a(S)$, simultaneously for all $S \subseteq B$, that is

$$\Pr(\forall S \subseteq B : \bar{a}(S) \leq a(S)) \geq 1-\alpha. \tag{3}$$

In ARI, the Simes-based critical vector is $l_i = i\alpha/h$, where h is a random variable that can be calculated using the short-cut defined by Goeman et al.[10] It is the largest set size of a subset of the brain not rejected by the Simes test.

The multiplicity control (3) that ARI guarantees is very versatile. It guarantees, simultaneously for every subset S of the brain, that the true activation $a(S)$ is at least as large as the claimed activation $\bar{a}(S)$. The analogous result for TDP follows immediately. Several more familiar error rates can be derived from Equation (3). Taking all clusters with TDP $= 1$ is equivalent to strong control of FWER at the voxel level. Taking clusters with TDP > 0 is equivalent to strong control of FWER at the cluster level but weak FWER control at the voxel level. At intermediate levels of TDP, Equation (3) gives intermediate information between weak and strong control at the voxel level. For more about relationships between Equation (3) and classical error rates, see Goeman et al.[18] Note that standard cluster-wise approaches based on the RFT or permutations[19,20] only provide strong control of the FWER at the cluster-level and weak control at the voxel-level, which is one of the error rates implied by Equation (3). The second feature, perhaps even more relevant, of Equation (3) is that the inference is simultaneous over all possible subsets of tested hypotheses (ie, voxels). Simultaneously implies that any exploratory and iterative approaches (ie, double-dipping) that are not possible in the cluster-wise approach become valid in the ARI class of methods. That is, the inferences on all subsets S are valid simultaneously and regardless of how they were selected (after seeing the data, changing the cluster-wise threshold, etc.).

Figure 1 illustrates computation of $\bar{a}(S)$ as defined in Equation (2), where $|S| = 1000$. In the left part, the length of the dashed black segments is the $1 - u + |\{i \in S : p_i \leq l_u\}|$ with $u \in \{1, \dots, |S|\}$ described in Equation (2), while the solid red segment is the maximum value over u, that is, the highest distance between the curve of observed p-values and critical vector $(l_1, \dots, l_m)$, for example, Simes-based. In the right part, we can see the trend of $\bar{a}(S)$ over u. The maximum value of $1 - u + z = 232$, where $z = |\{i \in S : p_i \leq l_u\}|$, is reached when u equals 97. This implies that $\bar{a}(S) = \max_{1\leq u\leq |S|} 1 - u + |\{i \in S : p_i \leq l_u\}| = 232$ is a lower confidence bound for the number of true discoveries in S.

The crucial assumption of Theorem 1 is that l_i satisfies Equation (1). In the case of the Simes test used by ARI, this follows from the PRDS assumption, commonly also adopted for the False Discovery Rate (FDR) controlling approach proposed by Benjamini et al.[21] Although this assumption is commonly accepted in neuroimaging,[1] the critical values $(l_1, \dots, l_m)$ can be overly strict if p-values are positively correlated, leading to conservative results. Moreover, the Simes critical vector may also be too strict or too loose if the p-values are not well calibrated.

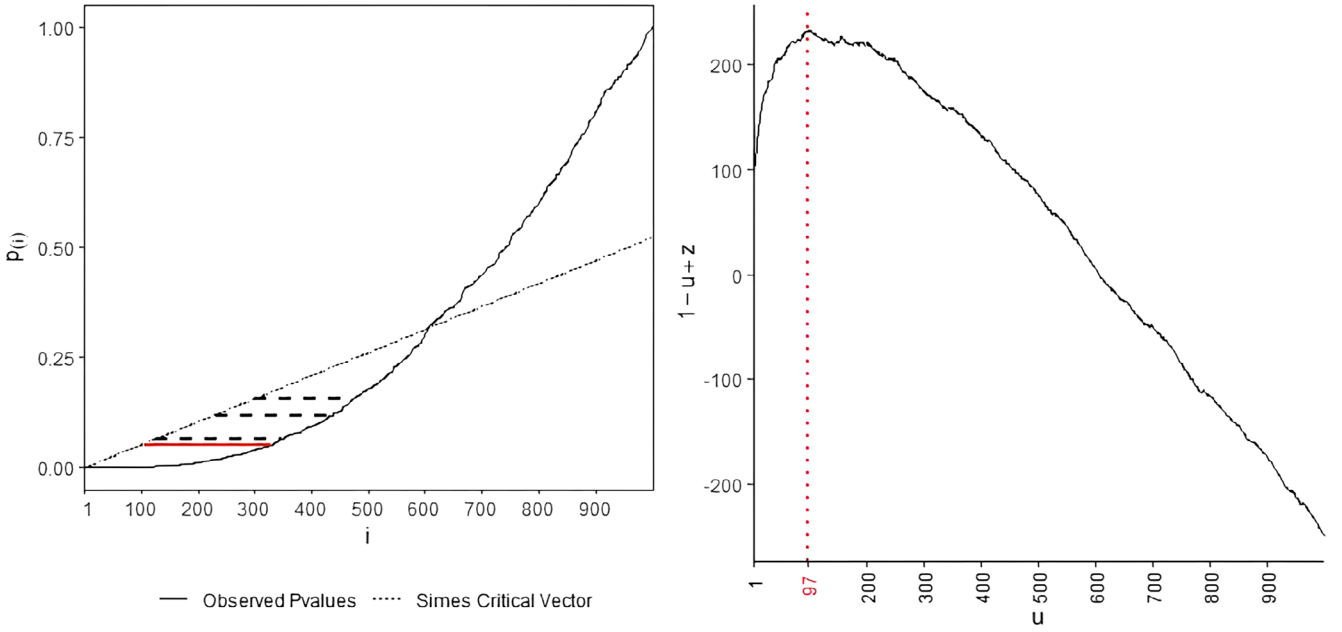

FIGURE 1 Left figure: A graphical display of the computation of $\bar{a}(S)$. The sorted p-values $p_{(i)}$ are plotted as solid line against their indexes i for $i = 1, \dots, 1000$. $\bar{a}(S)$ equals then the highest horizontal distance (solid red segment) between the curve of observed p-values $p_{(i)}$ and critical vector (dotted line) considering the distances represented by the black dashed long lines (here we put a sample). Right figure: Example of the computation of $\bar{a}(S)$, that is, the maximum $1 - u + z$ over u where $u \in \{1, \dots, 1000\}$ and $z = |\{i \in S : p_i \leq l_u\}|$. $\bar{a}(S)$ is then the maximum value of $1 - u + z$, represented by the red dotted line, attained when u equals 97.

3 | PERMUTATION-BASED ALL-RESOLUTIONS INFERENCE

To obtain a critical vector that leads to improved power, we propose a permutation procedure based on results in Hemerik et al.[14] The permutation method takes into account the dependence structure of the p-values and, therefore, often leads to a higher critical curve than parametric methods. Moreover, permutation methods not only adapt to the dependence structure but also to the marginal distributions of the p-values. This means that we do not require the null p-values to be uniformly distributed. Instead, we require that the null p-values are exchangeable with the corresponding post-permutation p-values (Assumption 1 in Hemerik et al).[14]

Following Hemerik et al,[14] we consider a group of permutations or sign-flipping transformations or any other data transformation that preserves the distribution of the test statistics under the null hypothesis, such as rotations.[22] These are maps from the support of the data distribution to itself. Our method is based on w random permutations or sign-flipping transformations. Let $p_1^1, ..., p_m^1 = p_1, ..., p_m$ be the p-values for the real data, and for every $2 \leq j \leq w$, let $p_1^j, ..., p_m^j$ be the p-values obtained for the j-th random permutation of the data.

Computing all possible permutations could be computationally infeasible, especially in the fMRI framework. However, Proposition 2 of Hemerik and Goeman[23] states that if the permutation set has a group structure, and $\alpha \in [0/w, 1/w, \dots (w-1)/w]$, the random permutations reach an exact α level. This means that the α level is exhausted if all hypotheses are true, and the error rate is at most α otherwise.

To obtain the permutation-based critical vector, the user must choose a family of candidate critical vectors. Examples of such candidate vectors are given in Section 5. We suppose that the candidate vectors are indexed by $\lambda_\alpha \in \Lambda \subseteq \mathbb{R}$, so that $l(\lambda_\alpha)$ denotes the candidate vector corresponding to λ_α. The family of candidate vectors is thus $\mathcal{F} = \{l(\lambda_\alpha) : \lambda_\alpha \in \Lambda\}$. We assume that the family of candidate vectors is monotone, in the sense that if $\lambda_\alpha^1, \lambda_\alpha^2 \in \Lambda$ and $\lambda_\alpha^1 \leq \lambda_\alpha^2$, then $l_i(\lambda_\alpha^1) \leq l_i(\lambda_\alpha^2)$ for every $1 \leq i \leq m$.

We define the permutation-based critical vector to be $l(\lambda_\alpha)$, where

$$\lambda_\alpha = \sup\{\lambda \in \Lambda : w^{-1}|\{1 \leq j \leq w : p_i^j \geq l_i(\lambda) \quad \forall i \in B\}| \geq 1 - \alpha\}. \tag{4}$$

By Hemerik et al,[14] the following holds, so that Theorem 1 applies.

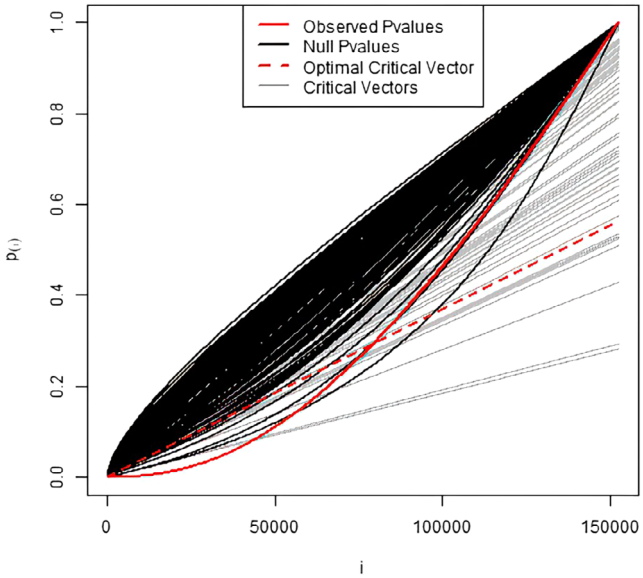

FIGURE 2 Example of $l(\lambda_\alpha)$ computation using $\alpha = 0.10$. The sorted p-values $p_{(i)}$ are plotted against their indexes $i = 1, \ldots, m$ with $m = 150{,}000$. The dashed red line represents the highest critical curve, that is, the optimal critical vector $(l_1(\lambda_\alpha), \ldots, l_m(\lambda_\alpha))$ than the gray ones, such that the $\alpha\%$ p-values distribution (black curves plus red one) is below it.

Theorem 2. *The vector $l(\lambda_\alpha)$ is a critical vector, that is, it satisfies (1) of Definition 1.*

The λ_α-calibration permits to incorporate the unknown dependence structure of the data into the choice of the critical vector. As seen from Equation (4), the λ_α value is computed in such a way that for at least $(1-\alpha)100\%$ of the permutations, all p-values lie above it. This is illustrated in Figure 2, considering a random sample of 100 permutation curves. The λ_α parameter tends to be lower if many null hypotheses are false, being εoptimalε if Equation (4) considers only $p_i \in B \setminus A$, that is, the set of true null hypotheses.

We use the critical vector $l(\lambda_\alpha) \in \mathcal{F}$ in Theorem 1 instead of the Simes-based one employed in ARI to gain power in computing $\bar{a}(S)$. The time complexity to compute the lower confidence bounds for the TDP is $\mathcal{O}(|S|log(|S|))$, after an initial step of calculating the critical values, which takes $\mathcal{O}(wmlog(m))$, using a similar algorithm as was used by Meijer et al[11] for the version with the parametric Simes test. So, it remains close to linear also if the whole brain is analyzed. Finally, the method can be uniformly improved by its iterative version,[14] presented in the next section.

4 | ITERATIVE APPROACH

We propose here an iterative method that uniformly improves $\bar{a}(S)$ defined in Equation (2) following the idea proposed by Hemerik et al.[14] The confidence envelope is defined as the minimum confidence bound computed in the complementary set of the rejection set having a cardinality equal to the lower bound of the number of true discoveries found in the previous iteration. The improvement is then substantial only when the number of detectable false hypotheses is large. In short, the iterative method improves $\bar{a}(S)$ sequentially in each step using the bound obtained in the previous step. The method always converges after a finite (and usually small) number of steps.

We rephrase below Theorem 2 of Hemerik et al[14] to get an improvement of the calibration parameter λ_α.

Theorem 3. *Let $\lambda_\alpha^0 = \lambda_\alpha$ as defined in Equation (4), we define $\lambda_\alpha(K)$ as,*

$$\lambda_\alpha(K) = \sup\{\lambda \in \Lambda : w^{-1}|\{1 \le j \le w : p_i^j \ge l_i(\lambda) \quad \forall i \in K\}| \ge 1-\alpha\}.$$

For $i \in \mathbb{N}$ and fixed $c \in [0,1]$ we consider $R = \{x \in B : p_x \le c\}$, and we determine:

$$\lambda_\alpha^{i+1} = \min\{\lambda_\alpha(K^c) : K \in R, |K| = \max_{1 \le u \le |R|} 1 - u + |\{i \in R : p_i \le l_u(\lambda_\alpha^i)\}|\}$$

where K^c is the complement of K. Then $\lambda_\alpha^0 \leq \lambda_\alpha^1 \leq \cdots$, and for a certain $i \in \mathbb{N}$, $\lambda_\alpha^i = \lambda_\alpha^{i+1}$. The function $l(\lambda_\alpha^{it})$ where $\lambda_\alpha^{it} = \max_{i \in \mathbb{N}} \lambda_\alpha^i$ is a critical vector in the sense of Definition (1).

Theorem (3) returns a confidence bound which uniformly improves the one defined in Theorem (1) with $l(\lambda_\alpha)$ defined in Theorem (2) as critical vector. Furthermore, if in every step we compute the improved bound for all $c \in [0, 1]$ and take the best one, then we are always better than the method proposed by Blanchard et al[17] if the same family of curves $\mathcal{F}$ is used. There are multiple versions of the iterative method, and that one (where in each step, we find the best c) is a uniform improvement of Blanchard et al[17] method.

In addition, we can simply demonstrate the uniform improvement over Blanchard et al[17] method by setting a specific $c \in [0, 1]$. First of all, by Definition (4), if $K_1 \subseteq K_2$, then $\lambda_\alpha(K_1) \geq \lambda_\alpha(K_2)$. Consequently $\lambda_\alpha(K_1)$ returns a greater confidence bound for $a(S)$ than the one calculated with $\lambda_\alpha(K_2)$. We can then focus on the size of the set K used to compute $\lambda_\alpha(K)$ to analyze the improvement of the iterative method. In the first step, the Blanchard et al[17] algorithm computes $\lambda_\alpha(K_1)$ with $|K_1| = |\{i \in B : p_i \geq l_1(\lambda_\alpha(B))\}| = k$. Instead, our iteration approach computes $\lambda_\alpha(K_2)$ where $|K_2| = m - \bar{a}(R)$. By definition of R, we can consider $c = l_1(\lambda_\alpha(B))$. Therefore, we have $|R| = m - k$ and $\bar{a}(R) \leq |R|$ which implies $|K_2| \leq m - m + k = k$, so $|K_2| \leq |K_1|$. This leads to $\lambda_\alpha(K_2) \geq \lambda_\alpha(K_1)$, and then we can say that the lower confidence bound proposed in Theorem 3 uniformly improves the one proposed by Blanchard et al.[17]

The iterative method is uniformly more powerful than the single-step method defined in Section 3, and also is uniformly more powerful than the step-down approach presented by Blanchard et al[17] under certain conditions. However, the power gain has as a cost a high computational time. In fact, the calculation of λ_α^{it} can be computationally infeasible if a large number of hypotheses is considered, as in the fMRI scenario. The iterative approach must compute the minimum of a set of size $|S|!/(|S| - \bar{a}^i(S))!\bar{a}^i(S)!$. Nevertheless, we suggest to use the approximated approach defined by Hemerik et al[14] which can be directly applied to our case. It simply calculates the minimum across sets randomly sampled from $\{K \subseteq S : |K| = \bar{a}^i(S)\}$ for $i \in \mathbb{N}$. The computation time, in this case, equals approximately 37 seconds analyzing 2000 hypotheses, 20 observations, and 1000 permutations. Finally, the approximated iterative approach provided valid inference in all simulations in Hemerik et al.[14] Please see Appendix D for further details.

5 | CHOICE OF FAMILY OF CURVES

In the previous section, we consider a general family $\mathcal{F}$ of candidate vectors $l(\lambda_\alpha)$, $\lambda_\alpha \in \Lambda$. Here we will discuss several examples of such families, which we considered in the application later in the paper.

The first family $\mathcal{F}$ that we consider is inspired by Simes' probability inequality.[9] The vectors are obtained by multiplying and shifting the Simes' critical vector. We denote the shift by $\delta \in \{0, \ldots, m-1\}$. For every such δ, we have a different family, indexed by $\lambda_\alpha \in \mathbb{R}$. The candidate critical vector $l(\lambda_\alpha)$ is defined by

$$l_i(\lambda_\alpha) = \frac{(i - \delta)\lambda_\alpha}{m - \delta}. \tag{5}$$

The shift parameter δ can be used to determine how sensitive the critical vector $l(\lambda_\alpha)$ will be to the smallest p-values. The parametric Simes-based approach corresponds to $\delta = 0$ and $\lambda_\alpha = \alpha$. We gain over that approach only if the λ_α value is greater than α.

Regarding the choice of δ, note that $l_i(\lambda_\alpha) \leq 0$ for $i \leq \delta$. As a consequence, we will find $\bar{a}(S) \leq |S| - \delta$, and $\bar{a}(S) = 0$ for all S with $|S| \leq \delta$. The value of δ, therefore, corresponds to the minimum size of a cluster that we are interested in detecting. To compensate, methods with large δ will often have a steeper slope λ_α and consequently have more power for detecting large clusters. In addition, the lower confidence bound $\bar{a}(S)$ computed by the shifted version, that is, $\delta > 0$, can not reach the 100% true discovery proportion, since the maximum equals to $(|S| - \delta)/|S|$. Further details about the shift parameter in computing bounds for the false discovery proportion can be found in Katsevich and Ramdas.[24]

The second example that we propose is a family $\mathcal{F}$ of candidate vectors that are derived from the asymptotically optimal rejection curves (AORC) considered in Finner et al[25] to control the FDR in an asymptotic Dirac uniform setting. Again we add a shift parameter $\delta \in \{0, \ldots, m\}$ as above, and we have a different family of candidate vectors for each δ. The calibration parameters lie in $\Lambda \subseteq \mathbb{R}$. The candidate critical vector $l(\lambda_\alpha)$ is defined by

$$l_i(\lambda_\alpha) = \frac{(i - \delta)\lambda_\alpha}{(m - \delta) - (i - \delta)(1 - \lambda_\alpha)}. \tag{6}$$

Our third example is related to the Higher Criticism method proposed by Donoho and Jin.[26] The candidate vectors are indexed by $\lambda_\alpha \in \Lambda \subseteq \mathbb{R}$ and are given by

$$l_i(\lambda_\alpha) = \frac{2i + \lambda_\alpha^2 - \sqrt{(2i + \lambda_\alpha^2)^2 - 4i^2(m + \lambda_\alpha^2)/m)}}{2(m + \lambda_\alpha^2)}. \tag{7}$$

Finally, we, we consider the family of candidate vectors $l(\lambda_\alpha)$ defined as follows:

$$l_i(\lambda_\alpha) = \inf\{x : \lambda_\alpha \leq F_i(x)\}. \tag{8}$$

Here $\lambda_\alpha \in \Lambda = [0, 1]$ and $F_i(X)$ is the cumulative distribution function of the beta distribution Beta$(i, m + 1 - i)$. This family was also considered in Hemerik et al.[14]

Further examples of candidate critical vectors can be found in references 17,27,28 and 14, but we did not consider them here. The results obtained with our permutation method will depend on the critical vector $l(\lambda_\alpha)$ and hence on the choice of the family $\mathcal{F} = \{l(\lambda_\alpha) : \lambda_\alpha \in \Lambda\}$. However, one choice of a family will essentially never lead to a uniform improvement compared to another family, but only to improved TDP bounds for some sets of hypotheses and worse bounds for other sets of hypotheses. Thus, the most appropriate family will depend on which set of hypotheses we are most interested in, for example, on whether we are interested in large or small clusters. Section 7 provides guidelines regarding a good choice of $\mathcal{F}$ for fMRI data.

6 | CHOICE OF PERMUTATIONS AND T-STATISTIC

In fMRI activation studies, the correlation between the (convolved) sequence of cognitive stimuli and the changes in BOLD response expresses brain activation. The changes in local hemodynamics affect the intensity of the magnetic resonance signal, that is, the voxel intensity. Therefore, the intensity of each voxel becomes the unit of interest. The differences in intensity, either between different conditions of an experiment or between different groups of participants, are expressed as a statistic value (t, z, or F usually), with an associated p-value. In fMRI, intensity values are often characterized by high spatial correlations and heteroscedasticity across voxels and across subjects due to the nature of the BOLD signal and external nuisance factors (eg, quality of the recording, respiration, or heartbeat). The permutation approach is useful in this situation, where parametric tests fail due to violations of assumptions.

In this section, we review which permutation test is valid and powerful to perform fMRI group analysis consisting of multi-subject studies to explore the differences in BOLD response recorded under two experimental conditions.[29-31] Group fMRI data are widely analyzed using a two-stage summary statistics approach within a mixed model. This approach uses ordinary least squares (OLS) methods,[32] in particular one- or two-sample t-tests, using within-subject parameter estimates as observations.

Let the first level within-subject model for each voxel $i \in \{1, \dots, m\}$ and each subject $j \in \{1, \dots, J\}$:

$$Y_{ij} = X_j \beta_{ij} + \epsilon_{ij},$$

where $Y_{ij} \in \mathbb{R}^n$ is the brain signal of subject j in voxel i, n is the total number of time points, J is the total number of subjects, m is the total number of voxels, $X_j \in \mathbb{R}^{n \times p}$ is the design matrix, where p regressors of interest, $\beta_{ij} \in \mathbb{R}^p$ is the vector of parameters, and $\epsilon_{ij} \in \mathbb{R}^n$ is the vector of autocorrelated and non-independent error terms. Let β_{1ij} be the parameter relative to the first experimental condition, while β_{2ij} to the second experimental condition for the subject j, we then assume for simplicity $p = 2$. We make inference on the contrasts of parameter estimates involving brain activation differences, that is, $D_{ij} = \hat{\beta}_{1ij} - \hat{\beta}_{2ij}$, so:

$$D_{ij} = \mu_i + \epsilon_{ij}^\star \tag{9}$$

where μ_i is the unknown parameter of interest representing the between-subject mean activation in voxel i, and $\epsilon_{ij}^\star$ are the error terms $\sim \mathcal{N}(0, \Sigma)$. To make inference on μ_i, the one-sample t-test is performed for each voxel i:

$$T_i = \frac{\hat{\mu}_i}{\sqrt{\hat{\sigma}_i{}^2/J}} \tag{10}$$

where $\hat{\mu}_i$ equals $\sum_{j=1}^{J} D_{ij}/J$ and $\hat{\sigma}_i^2$ equals $\sum_{j=1}^{J}(D_{ij}-\hat{\mu}_i)^2/(J-1)$. So, we have m statistical tests to analyze, one for each voxel i, that is, $H_{0i}: \mu_i = 0$, that create a statistical brain mapping.

Nevertheless, we need valid permutations to have a valid permutation testing procedure. It needs a null-invariant transformation of the data, that is, the joint distribution of the p-values under H_{0i} does not change.[23] In this case, $H_{0i}: \mu_i = 0$ implies that $(\beta_{1ij}, \beta_{2ij}) \overset{d}{=} (\beta_{2ij}, \beta_{1ij})$, that is equivalent to $D_{ij} \overset{d}{=} -D_{ij}$ for each voxel i. The compound symmetry is weaker than normality and also allows for heteroskedasticity. It can be justified by subtraction of two sample means with the same (arbitrary) distribution. Therefore, under H_{0i}, we can flip the sign at random of each D_{ij},[30] always taking the identity permutation as the first transformation to have an exact α method.[13,23]

The same approach can be used in the case of two-sample t-test. Let $G_j = \{1, 2\}$ expresses the group label for the j-th subject, the null hypothesis is then defined as $H_{0i}: \mu_{1i} = \mu_{2i}$. The exchangeability assumption implies $(D_{ij}|G_j = 1) \overset{d}{=} (D_{ij}|G_j = 2)$ for each voxel i, we can just shuffle the subject-group labels at random to compute the p-values null distribution. Permutation-based tests can be applied in various hypothesis testing's situations, for example, tests for linear models even in the presence of nuisance effects,[30,31,33] and tests for generalized linear models.[22]

7 | FMRI DATA APPLICATION

In this section, the permutation-based ARI method is evaluated using fMRI data. Two datasets from https://openneuro.org are analyzed. Both datasets have the same experimental design, that is, a block design with two stimuli. Pre-processing and first-level data analysis were performed using FMRIB Software Library (FSL).[34] Registration to Montreal Neurological Istitute (MNI) space was done using FMRIB's Linear Image Registration Tool (FLIRT),[34,35] motion correction using MCFLIRT,[36] and brain extraction using BET.[37] We applied spatial smoothing using a Gaussian kernel of 6mm full width at half minimum (FWHM). Finally, we applied a high-pass filter to the time-series data (Gaussian-weighted least-squares straight-line fitting, with sigma = 64.0 s). The parameter estimates (copes), that is, $\mathbf{D}_j \in \mathbb{R}^m$, were used as input in the `pARI`[38] package developed in `R`.[39] These parameter estimates are instead downloadable by installing the `fMRIdata R` package.[40]

For all the analyses, the α level is taken as 0.05 for a two-sided alternative hypothesis. We use 1000 permutations: 999 random permutations plus the identity. The approximated iterative approach (100 random combinations) presented in Section 4 is then applied. The results using the single-step method, that is, λ_α computed on the full set of hypotheses, are reported in Appendices A and B.

We chose δ as 0, 1, 9, and 27 to account for signal spreading out in clusters with size at least equals 0, 1, 9, and 27 voxels. The third powers were considered to exploit the three-dimensional structure of the voxels.

7.1 | Auditory data

We analyzed data from 140 subjects passively listening to vocal (ie, speech) and non-vocal sounds, collected by Pernet et al,[41] available at https://openneuro.org/datasets/ds000158/versions/1.0.0. We estimated the statistics map regarding the contrast that describes the difference of neural activation during vocal and non-vocal stimuli for each participant, that is, $\mathbf{D}_j$. The hypothesis testing is then constructed considering $H_{0i}: \mu_i = 0$ with two-sided alternative, where μ_i is the mean $\sum_{j=1}^{J} \hat{\boldsymbol{\beta}}_{\text{vocal }j} - \hat{\boldsymbol{\beta}}_{\text{non-vocal }j}/J$ computed for each voxel $i = 1, \ldots, m$, as described in Equation (10).

In concordance with results from earlier studies,[42-44] we found activation in the Frontal Pole (FP), Cingulate Gyrus (CG), Superior Frontal Gyrus (SFG), Temporal Occipital Fusiform Cortex (TOF), Lateral Occipital Cortex (LO), Lingual Gyrus (LG), Occipital Fusiform Gyrus (OFG), Inferior Temporal Gyrus (ITG), Supramarginal Gyrus (SG), Angular Gyrus (AG), Superior Temporal Gyrus (STG), Planum Temporale (PT), Middle Temporal Gyrus (MTG), Heschl's Gyrus (HG), Precentral Gyrus (PrG), Thalamus (T), Inferior Frontal Gyrus (IFG), Insular Cortex (I), Central Opercular Cortex (CO), and Frontal Medial Cortex (FM). While our method allows any method for forming clusters, we started from a map computed using RFT with a cluster-forming-threshold equalling $|T_i| > 3.2$. This threshold is quite liberal, and therefore we will make additional inferences inside these clusters with a threshold of $|T_i| > 4$.

Table 1 includes the lower bounds of the proportion of active voxels ($\overline{\pi}(S) = \bar{a}(S)/|S|$), the size of the cluster ($|S|$), the FWER-corrected p-values (p_{FWER}) from classical cluster analysis and the mm coordinates of the maximum. The FWER p-values based on the clusterwise RFT are reported only for the first cluster-forming-threshold equals to $|T_i| > 3.2$, since

TABLE 1 Auditory data: Clusters S identified with threshold $t = 3.2$ and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 1$) and parametric ARI, "drill down" clusters at $t = 4$.

Cluster	Threshold	Size	% active			RFT P-values	Voxel Coordinates		
S	t	$\lvert S\rvert$	$\overline{\pi}(S)$			p_{FWER}	x	y	z
			Perm Simes (11)	Perm AORC (12)	Parametric Simes (13)				
FP/CG/SFG/TOF/LO LG/OFG/ITG/SG/AG	3.2	40,094	96.77%	96.79%	84.98%	< 0.0001	−30	−34	−16
Left LO/TOF	4	8983	99.14%	99.14%	97.66%	–	−30	−34	−16
Right LO/LG/ITG	4	7653	98.96%	98.96%	97.25%	–	28	−30	−18
Left SFG/FP	4	1523	94.75%	94.81%	86.28%	–	−28	34	42
CG	4	1341	94.11%	94.11%	84.41%	–	6	40	−2
Right FP	4	1327	93.97%	93.97%	84.32%	–	30	56	28
Left SG/AG	4	859	90.69%	90.8%	75.79%	–	−50	−56	36
Right FP	4	243	67.08%	67.08%	43.21%	–	30	64	−4
Left SFG	4	202	61.88%	61.88%	40.1%	–	−18	8	52
Right SFG	4	122	46.72%	46.72%	19.67%	–	22	10	52
Right STG/PT/MTG HG/PrG/T	3.2	12,540	90.02%	90.05%	83.49%	< 0.0001	60	−10	0
STG/PT/MTG/HG	4	9533	99.19%	99.19%	97.8%	–	60	−10	0
PrG	4	485	86.19%	86.19%	78.35%	–	52	0	48
T	4	292	72.6%	72.6%	53.77%	–	10	−10	8
Left STG/PT/MTG/ HG/IFG/T	3.2	10,833	88.4%	88.45%	80.41%	< 0.0001	−60	−12	2
HG/PT/MTG/STG	4	7894	98.99%	98.99%	97.35%	–	−60	−12	2
IFG	4	667	88.01%	88.16%	74.06%	–	−40	14	26
T	4	34	26.47%	26.47%	17.65%	–	−14	−26	−4
Right IC/CO	3.2	408	37.25%	37.26%	24.01%	0.0002	38	−2	16
—	4	226	67.26%	67.26%	43.36%	–	38	−2	16
Left PrG	3.2	276	49.64%	49.64%	43.84%	0.002	−52	−6	50
—	4	192	71.35%	71.35%	63.02%	–	−52	−6	50
FM	3.2	270	22.59%	22.59%	13.33%	0.002	4	50	−14
—	4	128	47.66%	47.66%	28.13%	–	4	50	−14
SFG	3.2	187	6.95%	6.95%	0%	0.0123	6	52	38
—	4	64	20.31%	21.23%	0%	–	6	52	38
Left T	3.2	176	1.14%	1.14%	0%	0.0157	−14	−14	10
—	4	49	4.08%	4.08%	0%	–	−14	−14	10

Note: The size of the clusters $\lvert S\rvert$ and the voxel coordinates (x, y, z) are reported for each cluster.

this method does not allow double-dipping. The results are computed using the permutation-based ARI with the Simes and AORC family using $\delta = 1$. We compared these methods with the original parametric ARI calculated using the R package `ARIbrain`.[45]

As can be seen, the permutation-based ARI, that is, columns (11) for the Simes family and (12) for the AORC family in Table 1, has a better performance overall than the parametric approach, that is, column (13) in Table 1. However, the two families of candidate curves return very similar results; this likely reflects the similar structure of these two families of critical vectors. We also applied the shifted versions with $\delta > 1$ (see Appendix A for the results) but found that the loss of power in small clusters is not sufficiently offset by the gain in power in the larger clusters. We believe that this is due to the conservativeness of the null p-values, as shown in Figure 3. The family of critical vectors based on the Higher Criticism provide lower TDP than the ones given by the Simes and AORC families, and we put the results in Appendix A. The family based on the Beta quantile instead does not work on fMRI data due to the large number of variables that make the beta parameters unmanageable in terms of numerical precision. In addition, we believe that the weakness of both these two families is due also to mismatch between the design of the curves based on independent p-values that contrasts with a high correlation in the actual data.

Figure 4 shows the TDP bounds as a cluster brain map using the results using the Simes family confidence bound. In these maps, the user can directly interpret activation as the proportion of truly active voxels inside a cluster.

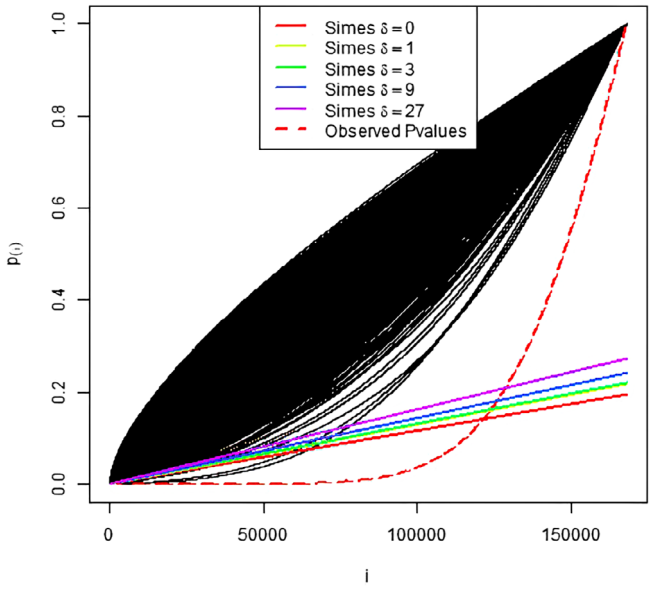

FIGURE 3 Auditory data: p-values null distribution (black lines plus dotted red one) with critical vectors from Simes family considering $\delta \in \{0, 1, 3, 9, 27\}$ (solid colored lines). The red dotted line represents the observed p-values.

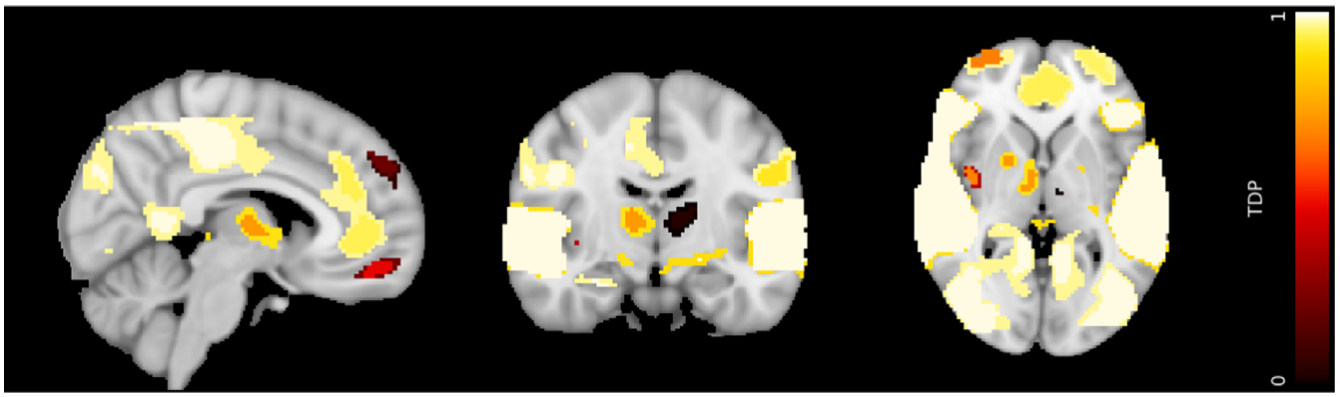

FIGURE 4 Auditory data: True discovery proportion map using the Simes family of critical vector with $\delta = 1$. Colors express the True Discovery Proportion for clusters based on a threshold of 3.2 and "drilled" down at 4.

7.2 | Rhyme data

Subsequently, we analyzed data from 13 subjects making rhyming judgments for pairs of either words or pseudo-words, collected by Xue and Poldrack[46] and available at https://openneuro.org/datasets/ds000003/versions/1.0.0. The analysis follows directly the one performed in Section 7.1, but the neural activation during the word stimulus was analyzed.

We found activity in Paracingulate Gyrus (PG), Lateral Occipital Cortex (LOC), Superior Frontal Gyrus (STG), Frontal Operculum Cortex (FOC), Putamen (P), Inferior Frontal Gyrus (IFG), Lingual Gyrus (LG), Occipital Fusiform Gyrus (OFG), Insular Cortex (IC), Cingulate Gyrus (CG), Superior Parietal Lobe (SPL), and Post Central Gyrus (PCG).[47] The cluster map is thresholded the same as the previous dataset: using cluster-wise RFT with a threshold of $|T_i| > 3.2$. We then drilled down $\overline{\pi}(S)$ using a threshold of $|T_i| > 4$. For completeness, we also analyzed the clusters defined by the threshold-free cluster enhancement (TFCE) method.[20] The analysis results are reported in Appendix B.

Figure 5 shows the null distribution of the p-values from the one-sample t-test (two-sided alternative) for the contrast regarding the word stimulus. As in Section 7.1, Table 2 represents the results using the Simes family in column

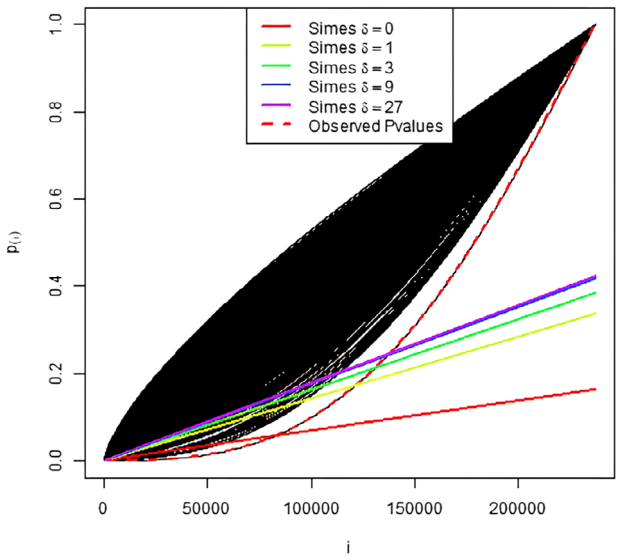

FIGURE 5 Rhyme data: p-values null distribution (black lines plus dotted red one) with critical vectors from Simes family considering $\delta \in \{0, 1, 3, 9, 27\}$ (solid colored lines). The red dotted line represents the observed p-values.

TABLE 2 Rhyme data: Clusters S identified with threshold $|\mathbf{T}| > 3.2$ and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 27$) and parametric ARI, "drill down" clusters at $|\mathbf{T}| > 4$.

Cluster	Threshold	Size	% active			RFT P-values	Voxel Coordinates		
S	t	$\lvert S\rvert$	$\overline{\pi}(S)$			p_{FWER}	x	y	z
			Perm Simes (14)	Perm AORC (15)	Parametric Simes (16)				
LOC/LG/OFG/PG/SFG	3.2	34,115	89.15%	89.4%	38.16%	< 0.001	4	12	48
FOC/P/IFG/IC/CG									
LOC/LG/OFG	4	11,045	91.21%	91.45%	42.01%	–	−6	−56	−12
FOC/P/IFG/IC	4	6930	85.75%	86.2%	29.32%	–	−42	14	−6
PG/SFG/CG	4	2100	57%	57.81%	18.05%	–	4	12	48
Left P	4	38	2.63%	2.63%	2.63%	–	−32	−18	−8
Left SPL/PCG	3.2	1546	1.49%	1.75%	0%	< 0.001	−24	−62	44

Note: The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

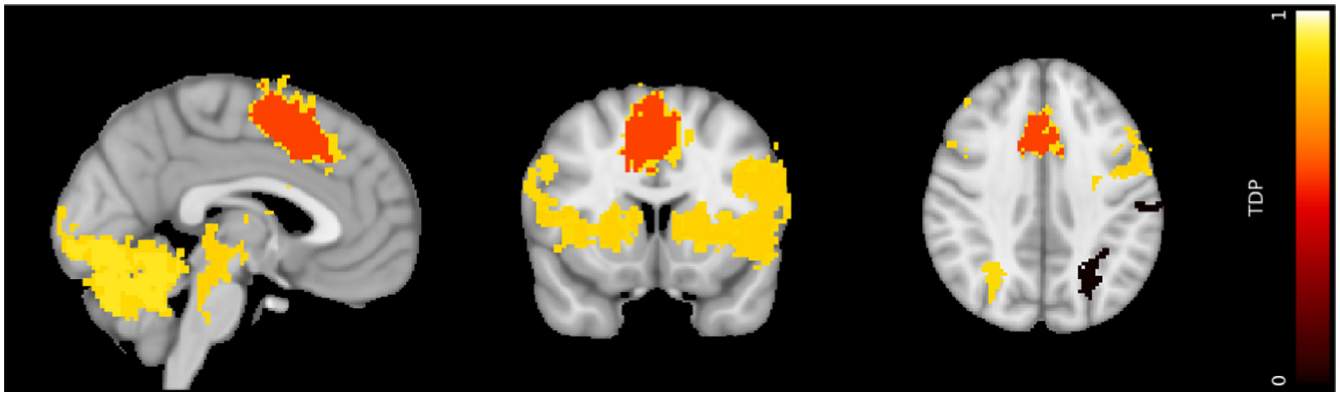

FIGURE 6 Rhyme data: True discovery proportion map using the Simes family of critical vector with $\delta = 27$. Colors express the true discovery proportion for clusters corresponding to a threshold of 3.2 and "drilled" down at 4.

(14) and the AORC family in column (15) with δ equalling 27. As can be seen, the power improvement over the parametric method, that is, column (16) of Table 2, is striking in this dataset. We decided to take δ equals 27 since, in this case, all clusters based on RFT have large sizes. The results using $\delta \in \{0, 1, 9\}$ are shown in Appendix B. Once again, the critical vectors based on Higher Criticism and the beta distribution do not work well due to the high correlation in the data.

Finally, Figure 6 shows the TDP represented in Table 2 as a cluster brain map.

8 | VALIDATING PERMUTATION-BASED ARI

FMRI data has noise characteristics that are hard to simulate using parametric distributions. Therefore, when performing simulations, often resting-state fMRI data (ie, fMRI data with no stimulus linked BOLD signal) is used. In these null data, the hypothesis of mean zero activation between groups is true while still retaining the noise characteristics of fMRI data. Eklund et al[3] found that many software programs, such as FSL[34] and Statistical Parametric Mapping (SPM),[48] do not properly control the probability or the average proportion of the false positives in cluster-wise inference when RFT assumptions are not met. As for ARI, RFT assumptions do not have to be met, we want to analyze the false positive rate of the permutation-based ARI using resting-state fMRI data with no signal. For this, we used the Oulu dataset provided by the 1000 Functional Connectomes Projects.[49] The pre-processing pipeline follows the one used in Eklund et al.[3] In particular, we analyzed the Oulu dataset from https://tinyurl.com/clusterfailure considering fMRI images pre-processed by FSL[34] with a level of smoothness equal to 6 mm FWHM and 6 different first level designs (four event activity paradigms, and two block activity paradigms).

The Oulu dataset consists of 103 subjects; however, to estimate the false positive rate, this set of subjects is not sufficient. In addition, Eklund et al[3] found asymmetric errors in the case of permutation test for the one-sample t-test using the Oulu dataset; therefore, we validate the permutation-based and parametric ARI, performing the two-sample t-tests. We select two groups of 20 subjects by randomly permuting 100 times the subject numbers and selecting the first 40 of this permuted dataset. Eklund et al[3] underline that the estimate of the familywise false positive rate is unbiased, even if these random datasets are not independent. Finally, the set of voxels used as a cluster map is used as the whole-brain mask. Please see Appendix G for the results of applying the one-sample t-tests.

Figure 7 shows the FWER estimated considering six different first level designs (ie, two-block activity paradigms: boxcar10 (10-s on-off), boxcar30 (30-s on-off) and four event activity paradigms: E1 (single event of 2-s activation, 6-s rest), E2 (single event 1- to 4-s activation, 3- to 6-s rest, randomized), E3 (13 events of 3–6 s for each task), and E4 (13 events of 3–6 s for each task, randomized). See references 3 and 50 for more details about tested parameter combinations.

To sum up, in most cases, the parametric-based ARI returns a false positive rate equal to 0, while the permutation-based ARI with the Simes family returns false positive rates greater than 0. Considering the boxcar10, E1, and E2 designs, the families with lower shifts, that is, 0 and 1, are more powerful than imposing the shift equals 3, 9, and 27. Therefore, both methods (ie, parametric-based and permutation-based ARI) control the FWER. In addition, the analysis confirms the conservativeness of the parametric ARI method in case of strong positive dependence due to the Simes inequality and positive regression dependence on subsets (PRDS)[12,17] assumptions.

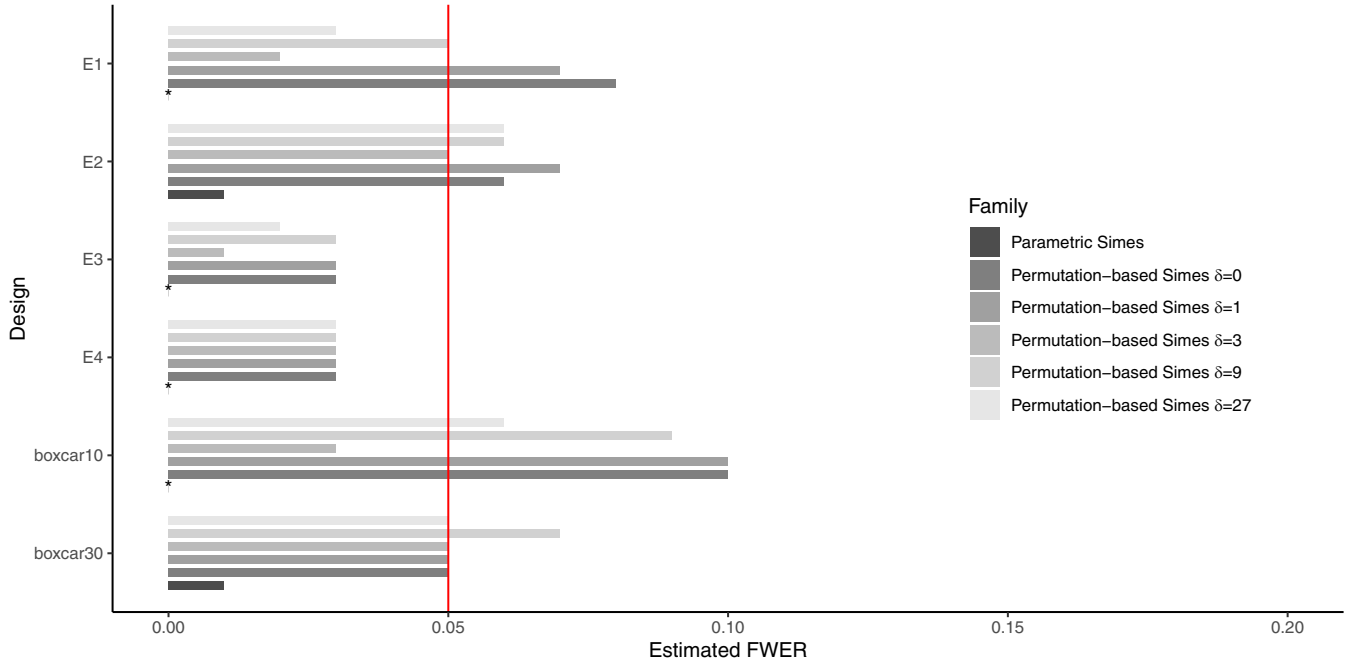

FIGURE 7 Estimated FWER considering six different first level designs, that is, two-block activity paradigms: boxcar10 (10-s on-off), boxcar30 (30-s on-off), and four event activity paradigms, that is, E1 (single event of 2-s activation, 6-s rest), E2 (single event 1- to 4-s activation, 3- to 6-s rest, randomized), E3 (13 events of 3–6 s for each task), and E4 (13 events of 3–6 s for each task, randomized), and six different methods to compute the TDP's lower bound (parametric Simes and permutation-based Simes considering five different values of the shift parameter, that is, $\delta \in \{0, 1, 3, 9, 27\}$). The solid red line represents the estimated nominal FWER equals 0.05, while the star symbols describe the estimated FWER equals 0.

9 | SIMULATION STUDY

We simulate data considering the simple following model (ie, model (9)):

$$D_{ij} = \mu_i + \epsilon_{ij}^{\star}$$

where $\mathbf{D}_j \in \mathbb{R}^m$, with $j = 1, \ldots, J$, J is the number of independent observations (ie, subjects) and m is the total number of voxels. The noise $\epsilon_j^{\star} \in \mathbb{R}^m$ follows the multivariate normal distribution with mean 0 and spatial correlation structure, that is, $\epsilon_j^{\star} \sim \mathcal{N}(0, \Sigma_\theta)$, where θ describes how rapidly the correlation declines with respect to the distance between two voxels. The three-dimensional coordinates of the voxels are defined as all combinations of vector $c = \{1, \ldots, m^{1/3}\}$, then $\Sigma_\theta = \exp(-\theta K)$ where K is the matrix containing the euclidean distances between the three-dimensional coordinates' voxels. For example, if $\theta = 0.2$, the correlation between two voxels with a distance of 1 equals 0.819, while the correlation between two voxels with a distance of 5 equals 0.368, and so on. The signal $\boldsymbol{\mu} \in \mathbb{R}^m$ is computed considering the difference in means having power of the one-sample t-test equals 0.8, that is, $\boldsymbol{\mu} = (z_{1-\alpha/2} + z_{1-\beta})/\sqrt{J}$, where $\alpha = 0.05$ is the significance level, $\beta = 0.8$ is the power level, and z_a is the quantiles of the standard normal distribution at level a. The signal $\boldsymbol{\mu}$ is equal to 0 under the null hypothesis.

First of all, we want to understand how the improvement of the nonparametric TDP lower bound changes concerning θ and the proportion of null hypotheses π_0. Let $J = 50$, $m = 1000$, $\theta \in \{0, 0.01, \ldots, 0.5\}$ and $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$, we simulate data 1000 times and the mean of $\overline{\pi}(S_m)$ over simulation is represented. The Simes family of confidence bound without shift is taken into account to compare with the parametric approach directly. Having no prior knowledge about the structure of the set of hypotheses to analyze, we consider the full set of hypotheses, that is, S_m. Figure 8 shows the difference of $\overline{\pi}(S_m)$ computed using the permutation and parametric methods over the θ and π_0 values. As expected, the permutation approach gets some power with respect to the parametric one in the case of correlation between pairs of variables. It can handle any type of dependence structure of the p-values.

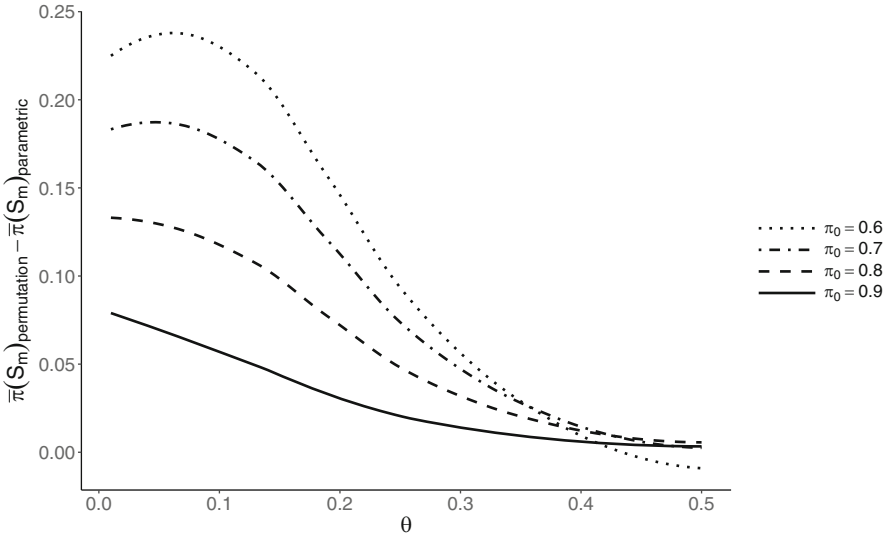

FIGURE 8 Difference of lower bounds for the true discoveries proportion considering the permutation $\overline{\pi}(S_m)_{\text{permutation}}$ and parametric $\overline{\pi}(S_m)_{\text{parametric}}$ methods using simulated data and considering the full set of hypotheses S_m over different values of $\theta \in \{0, 0.01, \ldots, 0.5\}$ and $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$.

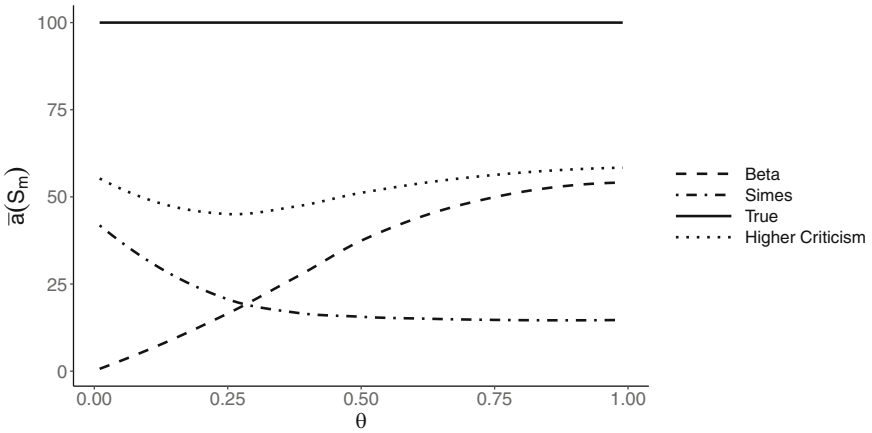

FIGURE 9 Simulated true discovery lower bound over S_m and different values of $\theta \in \{0, 0.01, \ldots, 0.99, 1\}$ using the Higher Criticism (dotted line), Beta (dashed line) and Simes critical vectors (dotted dashed line). The solid line represents the number of true discoveries, which equals 100 considering 1000 variables, and the proportion of null hypotheses $\pi_0 = 0.9$.

Secondly, we want to examine why certain families of critical curves do not provide good results in Section 7. The Higher Criticism critical vector (7), the Beta critical vector (8), and the Simes critical vector (5) are then used to compute $\bar{a}(S_m)$ using simulated data with $\pi_0 = 0.9$, $m = 1000$ and $J = 50$. As previously, we repeat the simulations 1000 times for each framework, and the mean value of $\bar{a}(S_m)$ is computed. Figure 9 shows the behavior of these three families of critical vectors with respect to $\theta \in \{0, 0.01, \ldots, 0.99, 1\}$. In Section 5, we said that the Higher Criticism and Beta families could be problematic in the case of a strong correlation between tests. As expected, the Beta critical vector does not work in the case of a strong correlation between variables, that is, low values of θ. However, the Higher Criticism family seems to work considering various values of θ in contrast to the results with fMRI data. This may be due to the different spatial correlation in the fMRI data, which is much more complex than that specified in the simulations. However, it can be seen that the lower bound for the TDP calculated by the Higher Criticism family is close to the one computed by the Simes family as the correlation increases. Finally, the Beta family works only if $\theta > 0.2$ (ie, correlation between voxels equals 0.28 on average). However, high values of correlation are unrealistic in real applications. For example, the mean correlation across 10,000 randomly sampled voxels equals 0.25 in the case of Rhyme data.

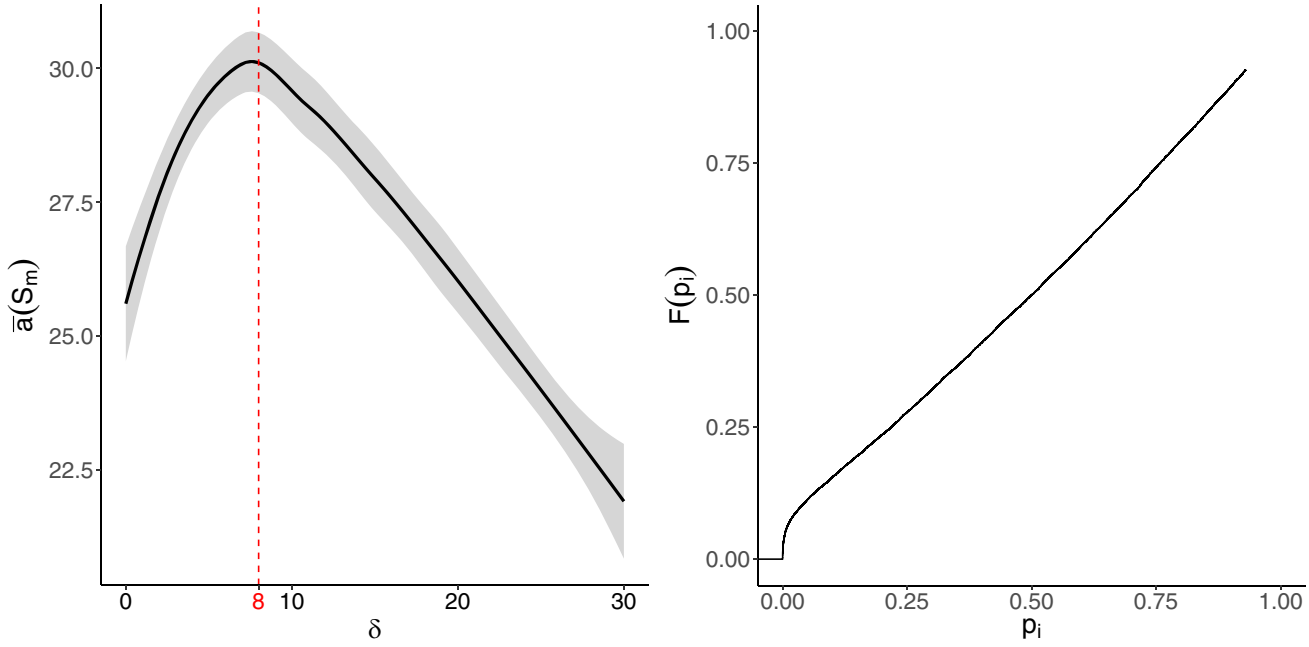

FIGURE 10 Left side: True discovery lower bound using simulated data. The full set of hypotheses S_m is considered over different values of δ. Right side: Empirical cumulative density function of observed raw p-values, that is, $F(p_i)$.

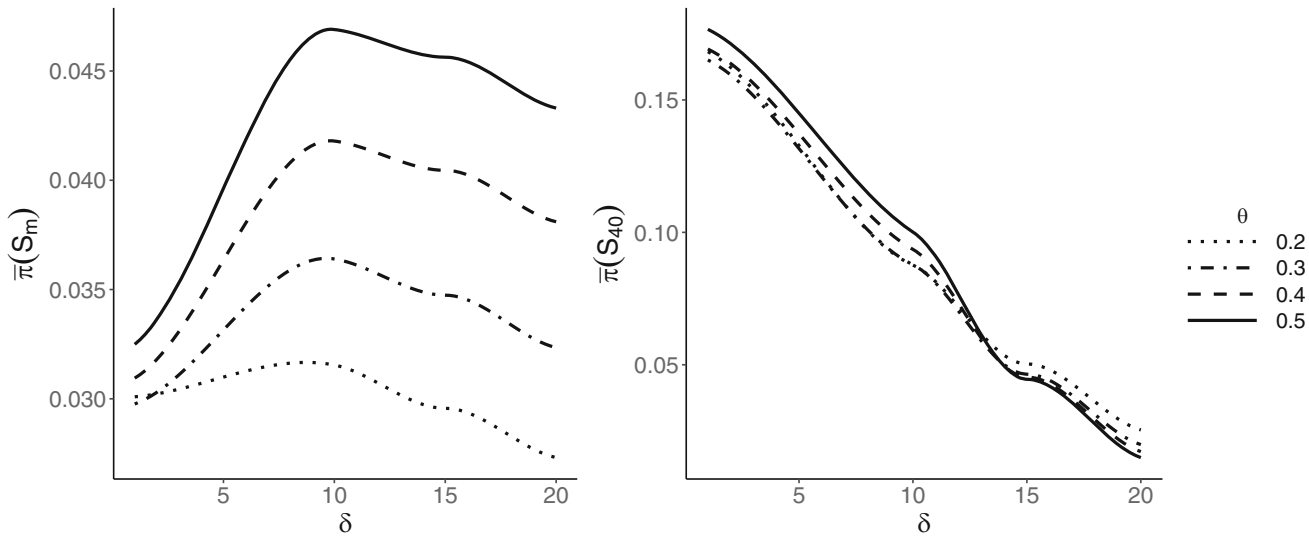

FIGURE 11 Lower Bounds for the true discovery proportion using simulated data with $\theta \in \{0.2, 0.3, 0.4, 0.5\}$. In the left figure, the full set of hypotheses is considered, while in the right figure a random sample of 40 hypotheses is analyzed. The critical vectors based on the Simes family with $\delta \in \{0, 5, 10, 15, 20\}$ are used in both situations.

Thirdly, we want to analyze how the Simes family of critical curves (5) works if anti-conservative p-values distribution is considered. Let $J = 50$, $m = 1000$, $\theta = 0.2$ (ie, correlation between voxels equals 0.28 on average), and $\pi_0 = 0.9$, we compute $\bar{a}(S_m)$ for every 1000 simulations, and once again the mean over simulations is reported. Figure 10 shows $\bar{a}(S_m)$ considering the Simes family using $\delta \in \{0, \ldots, 30\}$. We can note that the shifted version works well in the case of anti-conservative p-values if the corrected value for the tuning parameter δ is chosen, described by the red dotted line, that is, $\delta = 8$.

Therefore, we explore how the Simes family of critical curves (5) works with different values of θ and S size. The left part of Figure 11 shows the mean of the lower bounds for the true discoveries considering the full set of hypotheses,

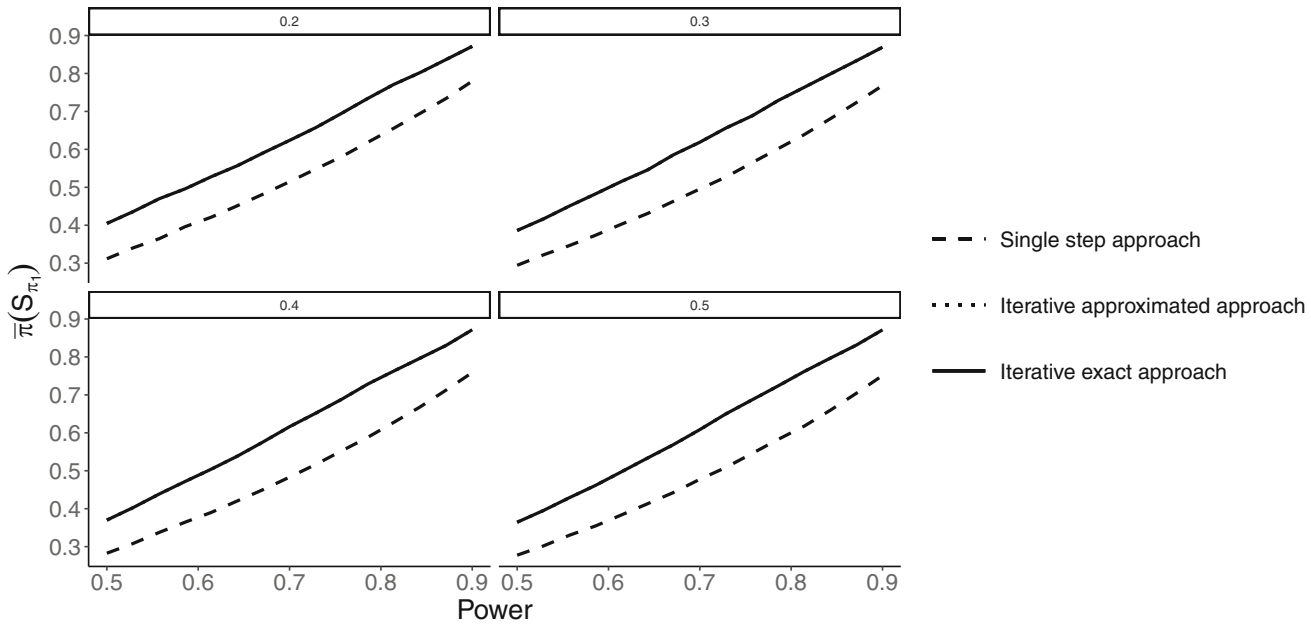

FIGURE 12 Simulated true discovery proportion lower bound for S_{π_1} over different values of θ and power using the single-step (dashed line), iterative approximated version (dotted line) and iterative exact version (solid line). The dotted line is behind the solid line.

that is, S_m, and $\delta \in \{0, 5, 10, 15, 20\}$ over 1000 simulations. We can note that in almost all scenarios, the shifted version outperforms the unshifted ones. The difference gets smaller if θ decreases (ie, the correlations between voxels increase). However, the situation changes if we compute the TDP for a smaller set of hypotheses than S_m as shown in the right part of Figure 11. In this case, we randomly sample 40 hypotheses from the false null ones, that is, S_{40}.

Finally, the performance of the iterative approach proposed in Section 4 is compared with respect to the single-step one presented in Section 3. Let consider directly S_{π_1} the set of true discoveries, therefore $\pi(S_{\pi_1}) = 1$. Figure 12 shows the true discovery proportion $\overline{\pi}(S_{\pi_1})$ computed on 1000 simulated data with π_0 equals 0.9, $\theta \in \{0.2, 0.3, 0.4, 0.5\}$ and different levels of power used to simulate the data. In this case, we consider $m = 64$ so that we can use the exact iterative method. First of all, we can see how the approximated iterative version equals the exact one and, more importantly, how both of them uniformly improve the single-step approach.

To sum up, we suggest using the Higher Criticism and Beta families if the correlation across the variables is supposed to be low. Besides, we recommend considering the shifted version of the Simes or AORC families if the interest is in large sets of hypotheses rather than small ones. The shifted version of the Simes and AORC families is again recommended if the p-values' distribution is expected to be anti-conservative. We stress that the value of the δ parameter must be decided a priori and chosen reasonably concerning the data analyzed, as seen in the fMRI data application.

We include some simulation analyses to examine the power of the iterative approach presented in Section 4 and the influence of the number of combinations chosen for its approximated version in Appendix D. Finally, we show some simulation studies in Appendix H in the case of equi-correlation variance structure for $\epsilon_j^\star$.

10 | CONCLUSIONS

Our proposed method finds simultaneous lower bounds for the TDP over all possible hypothesis subsets using the permutation theory in a computationally efficient way. As a simultaneous method, it allows the decision of which hypothesis sets to analyze to be entirely flexible and post-hoc, that is, the user can choose it after seeing the data and revise the choice as often as he/she wants. It is particularly useful in fMRI single and multi-subjects analysis to infer inside clusters, resolving the so-called spatial specificity paradox, without falling into the double-dipping problem.

A method that has some apparent similarity with the permutation-based ARI approach is the TFCE approach proposed by Smith and Nichols.[20] Both methods use permutation theory, and both are flexible in the choice of threshold. There are two important differences between the methods, however. First, while TFCE allows data-driven thresholds, our proposed method is more flexible since it allows the simultaneous use of many thresholds, which can be chosen after viewing the data. Second—and more importantly—permutation-based ARI provides additional information about the clusters: a lower bound for the proportion of true discoveries. In contrast, TFCE remains a cluster-level inferential method, returning only a p-value for the clusterwise null hypothesis.

Permutation-based methods are recommended whenever they can be used, gaining power over the parametric approaches, especially when the p-values are strongly dependent, as for fMRI data. In this work, we used permutation theory to calculate the critical vector needed for ARI. Our method adapts to the correlation structure of the data in an exact way by means of the calibration of the parameter λ_α. In this way, our method remedies the existing issues of anti-conservativeness and conservativeness. Indeed, we found that the permutation-based method has more power than the parametric approach both in simulated and real data and confirmed FWER control using resting-state null data.

Permutation methods are not assumption-free but require the exchangeability of the test statistics under the null hypothesis. We showed the results using the OLS one-sample t-test for fMRI group analysis, having a fast and straightforward computation of the permutation null distribution, randomly flipping the sign of each subject's contrast. The exchangeability assumption needed to perform permutation-based methods is satisfied, that is, the error terms of the model need to be symmetric around 0. However, the method is also applicable using other statistical tests, for example, two-sample t-tests. Even if permutations are employed to perform the method proposed, the computation time remains low, for example, around 210 s using the single-step method, while around 1 h using the iterative approach with 10 combinations, having 150,000 hypotheses and 1000 permutations. The computation time is related to a device with a processor having 1.8 GHz CPU and 16 GB of RAM, finally, the R package `pARI` available on CRAN[38] based on the C++ language was used.

The proposed method is general, allowing different families of confidence bounds. The choice of the family is critical since it directly influences the bounds for the true discovery proportions and, thus, the power properties of the method. Simulations and real data analysis suggest the Simes (5) and AORC (6) families in the fMRI framework, while the Higher Criticism (1) and Beta families (8) if the correlation between variables is supposed to be low. Simes (5) and AORC (6) families depend on the shift parameter δ. We recommend fixing $\delta = 1$ if the practitioner is interested in computing the lower bound for the TDP in small clusters, while $\delta > 1$ if the attention is focused on large clusters. Finally, drilling down may increase or decrease the lower bound for the TDP of some subclusters if a large cluster is analyzed, for example, the first cluster found in Section 7.2. This suggestion is also confirmed by the simulation analysis. We found that the shifted versions gain power if the raw p-values are anti-conservative. Other types of families that we analyzed, based on Higher Criticism[26] and Beta quantiles, do not seem to perform well in fMRI data analysis due to the strong correlation among the voxels, as also illustrated in the simulation study of Section 9. Generally, we suggest a family of critical vectors more concentrated on small p-values if the number of rejected hypotheses may be low and a family of critical vectors more diffuse if the number of rejected hypotheses may be high.

Finally, we implement the iterative approach proposed by Hemerik et al[14] in the simultaneous post-hoc inference scenario, which uniformly improves the Blanchard et al[17] bounds in most cases. There is a power gain here, but for fMRI data with sparse signal, the gain is small and comes at a large computational cost.

Our presently used method provides a useful and practical selective inference for fMRI data that exploits the advantages of permutation theory and the closed-testing procedure, resolving the spatial specificity paradox with quite fast computation time. The proposed method would be applicable not only for the fMRI data but more for any other data types that may yield multiple testing problems and cluster-wise inference (eg, electroencephalography data and genomic data).

AUTHOR CONTRIBUTIONS

All authors have directly participated in the planning and execution of the presented work.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The fMRI data used are available from http://github.com/angeella/fMRIdata, and the approach is developed as R package available on CRAN (https://CRAN.R-project.org/package=pARI). The Oulu dataset is available at https://tinyurl.com/clusterfailure.

ORCID

Angela Andreella https://orcid.org/0000-0002-1141-3041

REFERENCES

1. Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res.* 2003;12(5):419-446.
2. Worsley KJ, Sean M, Peter N, Vandal Alain AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp.* 1996;4(1):58-73.
3. Eklund A, Nichols ET, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA.* 2016;113(28):7900-7905.
4. Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage.* 2014;91:412-419.
5. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience – the dangers of double dipping. *Nat Neurosci.* 2009;12:535-540.
6. Rosenblatt JD, Finos L, Wouter DW, Solari A, Goeman JJ. All-resolutions inference for brain imaging. *Neuroimage.* 2018;181:786-796.
7. Goeman JJ, Solari A. Multiple testing for exploratory research. *Stat Sci.* 2011;26(4):584-597.
8. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63(3):655-660.
9. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika.* 1986;73(3):751-754.
10. Goeman JJ, Meijer RJ, Krebs TJP, Solari A. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika.* 2019;106(4):841-856.
11. Meijer RJ, Thijmen JP, Krebs TJP, Goeman JJ. Hommel's procedure in linear time. *Biom J.* 2019;61:73-78.
12. Sarkar SK. On the Simes inequality and its generalization. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. 2008;1:231-242.
13. Pesarin F, Salmaso L. *Permutation Tests for Complex Data: Theory, Applications and Software*. New York: John Wiley and Sons; 2010.
14. Hemerik J, Solari A, Goeman JJ. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika.* 2019;106(3):635-649.
15. Winkler AM, Ridgway GR, Douaud G, Nichols TE, Smith SM. Faster permutation inference in brain imaging. *Neuroimage.* 2016;141:502-516.
16. Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE. Non-parametric combination and related permutation tests for neuroimaging. *Hum Brain Mapp.* 2016;37:1486-1511.
17. Blanchard G, Neuvial P, Roquain E. Post hoc confidence bounds on false positives using reference families. *Ann Stat.* 2020;48(3):1281-1303.
18. Goeman JJ, Jesse H, Aldo S. Only closed testing procedures are admissible for controlling false discovery proportions. *Ann Stat.* 2021;49(2):1218-1238.
19. Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp.* 2002;15(1):1-25.
20. Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage.* 2009;44(1):83-98.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc: Ser B (Stat Methodol).* 1995;57(1):289-300.
22. Hemerik J, Goeman JJ, Finos L. Robust testing in generalized linear models by sign flipping score contributions. *J R Stat Soc: Ser B (Stat Methodol).* 2020;82(3):841-864.
23. Hemerik J, Goeman JJ. Exact testing with random permutation. *TEST.* 2018;27(4):811-825.
24. Katsevich E, Ramdas A. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *Ann Stat.* 2020;48(6):3465-3487.
25. Finner H, Dickhaus T, Roters M. On the false discovery rate and an asymptotically optimal rejection curve. *Ann Stat.* 2009;37(2):596-618.
26. Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat.* 2004;32(3):962-994.
27. Durand G, Blanchard G, Neuvial P, Roquain E. Post hoc false positive control for structured hypotheses. *Scand J Stat.* 2020;47(4):1114-1148.
28. Blanchard G, Roquain E. Two simple sufficient conditions for FDR control. *Electron J Stat.* 2008;2:963-992.

29. Holmes AP, Blair RC, Watson JDG, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. *J Cereb Blood Flow Metab.* 1996;16(1):7-22.
30. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. *Neuroimage.* 2014;92:381-397.
31. Helwig NE. Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *Wiley Interdiscip Rev: Comput Stat.* 2019;11(2):e1457.
32. Mumford JA, Nichols T. Simple group fMRI modeling and inference. *Neuroimage.* 2009;47(4):1469-1475.
33. Solari A, Finos L, Goeman JJ. Rotation-based multiple testing in the multivariate linear model. *Biometrics.* 2014;70(4):954-961.
34. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62(2):782-790.
35. Jenkinson M, Smith SM. A global optimization method for robust affine registration of brain images. *Med Image Anal.* 2001;5(2):143-156.
36. Jenkinson M, Bannister P, Brady M, Smith SM. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage.* 2002;17(2):825-841.
37. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp.* 2002;17(3):143-155.
38. Andreella A. pARI: Permutation-Based All-Resolutions Inference Method. R package version 1.1.1. 2022. https://CRAN.R-project.org/package=pARI
39. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017.
40. Andreella A. fMRIdata: Preprocessed fMRI data from. 2020. https://openneuro.org/ https://github.com/angeella/fMRIdata
41. Pernet CR, Belin P, McAleer P, et al. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *OpenNeuro.* 2019; https://openneuro.org/datasets/ds000158/versions/1.0.0
42. Pernet CR, McAleer P, Latinus M, et al. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage.* 2014;119:164-174.
43. Olson IR, Gatenby JC, Gore JC. A comparison of bound and unbound audio–visual information processing in the human cerebral cortex. *Brain Res.* 2002;14(1):129-138.
44. Calvert GA, Bullmore ET, Brammer MJ, et al. Activation of auditory cortex during silent lipreading. *Science.* 1997;276(5312):593-596.
45. Finos L, Goeman JJ, Weeda W, Rosenblatt J, Solari A. ARIbrain: All-Resolutions Inference. R package version 0.2. 2018. https://CRAN.R-project.org/package=ARIbrain
46. Xue G, Poldrack RA. Rhyme judgment. *OpenNeuro.* 2020; https://openneuro.org/datasets/ds000003/versions/1.0.0
47. Xue G, Poldrack RA. The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. *J Cogn Neurosci.* 2007;19(10):1643-1655.
48. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Edinburgh: Elsevier; 2011.
49. Biswal BB, Mennes M, Zuo XN, et al. Toward discovery science of human brain function. *Proc Natl Acad Sci USA*. 2010;107(10):4734-4739.
50. Anders E, Hans K, Nichols TE. Cluster failure revisited: impact of first level design and physiological noise on cluster false positive rates. *Hum Brain Mapp.* 2019;40(7):2017-2032.
51. Duncan K, Pattamadilok C, Knierim I, Devlin J. Word and object processing. *Stanford Digital Repository*. 2009 http://purl.stanford.edu/nb256hg3654 and https://openfmri.org/dataset/ds000107/

How to cite this article: Andreella A, Hemerik J, Finos L, Weeda W, Goeman J. Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*. 2023;1-30. doi: 10.1002/sim.9725

APPENDIX A. AUDITORY DATA ANALYSIS

Table A1 contains the results using the single-step approach presented in Section 3 with $\delta = 1$ as Table 1 of Subsection 7.1. Instead, Table A2 shows the results using the Simes family of confidence bounds, considering δ equal respectively to 0, 9, and 27. In the same way, Table A3 proposes the results using the AORC family. Finally, Table A4 shows the results using the family of critical vectors based on the Higher Criticism. In all analyses, we consider α equals 0.05.

TABLE A1 Auditory data: Clusters S identified with threshold $t = 3.2$ and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 1$) and parametric ARI, "drill down" clusters at $t = 4$.

Cluster	Threshold	Size	% active			RFT P-values	Voxel Coordinates		
S	t	$\lvert S\rvert$	$\overline{\pi}(S)$			p_{FWER}	x	y	z
			Perm Simes (A1)	Perm AORC (A2)	Parametric Simes (A3)				
FP/CG/SFG/TOF/LO LG/OFG/ITG/SG/AG	3.2	40,094	96.7%	96.73%	84.98%	< 0.0001	−30	−34	−16
Left LO/TOF	4	8983	99.11%	99.11%	97.66%	–	−30	−34	−16
Right LO/LG/ITG	4	7653	98.96%	98.96%	97.25%	–	28	−30	−18
Left SFG/FP	4	1523	94.75%	94.75%	86.28%	–	−28	34	42
CG	4	1341	94.11%	94.11%	84.41%	–	6	40	−2
Right FP	4	1327	93.97%	93.97%	84.32%	–	30	56	28
Left SG/AG	4	859	90.69%	90.8%	75.79%	–	−50	−56	36
Right FP	4	243	67.08%	67.08%	43.21%	–	30	64	−4
Left SFG	4	202	61.88%	61.88%	40.1%	–	−18	8	52
Right SFG	4	122	46.72%	46.72%	19.67%	–	22	10	52
Right STG/PT/MTG HG/PrG/T	3.2	12,540	89.82%	89.86%	83.49%	< 0.0001	60	−10	0
STG/PT/MTG/HG	4	9533	99.16%	99.16%	97.8%	–	60	−10	0
PrG	4	485	86.19%	86.19%	78.35%	–	52	0	48
T	4	292	72.6%	72.6%	53.77%	–	10	−10	8
Left STG/PT/MTG/ HG/IFG/T	3.2	10,833	88.4%	88.45%	80.41%	< 0.0001	−60	−12	2
HG/PT/MTG/STG	4	7894	98.99%	98.99%	97.35%	–	−60	−12	2
IFG	4	667	88.01%	88.19%	74.06%	–	−40	14	26
T	4	34	26.47%	26.47%	17.65%	–	−14	−26	−4
Right IC/CO	3.2	408	37.25%	37.25%	24.01%	0.0002	38	−2	16
—	4	226	67.26%	67.26%	43.36%	–	38	−2	16
Left PrG	3.2	276	49.63%	49.64%	43.84%	0.002	−52	−6	50
—	4	192	71.35%	71.35%	63.02%	–	−52	−6	50
FM	3.2	270	22.59%	22.59%	13.33%	0.002	4	50	−14
—	4	128	47.66%	47.66%	28.13%	–	4	50	−14
SFG	3.2	187	6.95%	6.95%	0%	0.0123	6	52	38
—	4	64	20.31%	21.23%	0%	–	6	52	38
Left T	3.2	176	1.14%	1.14%	0%	0.0157	−14	−14	10
—	4	49	4.08%	4.08%	0%	–	−14	−14	10

Note: The single-step method is applied. The size of the clusters $\lvert S\rvert$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE A2 Auditory data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using Simes family of critical vectors with $\delta \in \{0, 9, 27\}$.

Cluster	Size	% active			voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$			x	y	z
		$\delta = 0$	$\delta = 9$	$\delta = 27$			
FP/CG/SFG/TOF/LO/LG/OFG/ITG/SG/AG	40,094	96.33%	97.01%	97.31%	−30	−34	−16
Right STG/PT/MTG/HG/PrG/T	12,540	89.09%	90.56%	91.43%	60	−10	0
Left STG/PT/MTG/HG/IFG/T	10,833	87.33%	89.25%	90.1%	−60	−12	2
Right IC/CO	408	36.03%	36.76%	33.57%	38	−2	16
Left PrG	276	49.28%	47.46%	42.03%	−52	−6	50
FM	270	21.11%	21.11%	16.67%	4	50	−14
SFG	187	6.42%	3.74%	0%	6	52	38
Left T	176	1.14%	0%	0%	−14	−14	10

Note: The iterative method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE A3 Auditory data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using AORC family of critical vectors with $\delta \in \{0, 9, 27\}$.

Cluster	Size	% active			voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$			x	y	z
		$\delta = 0$	$\delta = 9$	$\delta = 27$			
FP/CG/SFG/TOF/LO/LG/OFG/ITG/SG/AG	40,094	96.37%	97.04%	97.33%	−30	−34	−16
Right STG/PT/MTG/HG/PrG/T	12,540	89.15%	90.64%	91.5%	60	−10	0
Left STG/PT/MTG/HG/IFG/T	10,833	87.42%	89.34%	90.16%	−60	−12	2
Right IC/CO	408	36.28%	36.76%	33.58%	38	−2	16
Left PrG	276	49.28%	47.46%	42.03%	−52	−6	50
FM	270	21.11%	21.11%	16.67%	4	50	−14
SFG	187	6.42%	3.74%	0%	6	52	38
Left T	176	1.14%	0%	0%	−14	−14	10

Note: The iterative method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE A4 Auditory data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using the family of critical vectors based on the higher criticism.

Cluster	Size	% active	voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$	x	y	z
		Higher Criticism			
FP/CG/SFG/TOF/LO/LG/OFG/ITG/SG/AG	40,094	95.66%	−30	−34	−16
Right STG/PT/MTG/HG/PrG/T	12,540	86.17%	60	-10	0
Left STG/PT/MTG/HG/IFG/T	10,833	83.99%	-60	-12	2
Right IC/CO	408	0%	38	-2	16
Left PrG	276	24.64%	-52	-6	50
FM	270	0%	4	50	−14
SFG	187	0%	6	52	38
Left T	176	0%	-14	-14	10

Note: The single-step method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

APPENDIX B. RHYME DATA ANALYSIS

Table B1 includes the results applying the single-step method presented in Section 3 imposing $\delta = 27$, following the structure of Table 2 of Subsection 7.2. The performance having $\delta \in \{0, 1, 9\}$ is proposed in Table B2 using the Simes family of confidence bounds and in Table B3 using the AORC family. Table B4shows the results using the family of critical vectors based on the Higher Criticism. In all the analyses, the α level equals 0.05. Finally, Table B5 shows the lower bounds for the TDP as before but considering clusters computed by threshold-free cluster enhancement (TFCE) method[20] using p-value threshold equals 0.05. We found activity in Lingual Gyrus (LG), Occipital Pole (OP), Putamen (P), Superior Frontal Gyrus (SFG), Frontal Pole (FP), Insular Cortex (I), Occipital Fusiform Gyrus (OFG), Lateral Occipital Cortex (LO), Precentral Gyrus (PrG), Post Central Gyrus (PCG), and Paracingulate Gyrus (PG).

TABLE B1 Rhyme data: Clusters S identified with threshold $t = 3.2$ and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 27$) and parametric ARI, "drill down" clusters at $t = 4$.

Cluster	Threshold	Size	% active			RFT	Voxel		
						P-values	Coordinates		
S	t	$\lvert S\rvert$	$\overline{\pi}(S)$			p_{FWER}	x	y	z
			Perm Simes (B4)	Perm AORC (B5)	Parametric Simes (B6)				
LOC/LG/OFG/PG/SFG	3.2	34,115	87.38%	87.85%	38.16%	< 0.001	4	12	48
FOC/P/IFG/IC/CG									
LOC/LG/OFG	4	11,045	90.82%	91.09%	42.01%	–	−6	−56	−12
FOC/P/IFG/IC	4	6930	85.38%	85.81%	29.32%	–	−42	14	−6
PG/SFG/CG	4	2100	56.95%	57.67%	18.05%	–	4	12	48
Left P	4	38	2.63%	2.63%	2.63%	–	−32	−18	−8
Left SPL/PCG	3.2	1546	1.49%	1.75%	0%	< 0.001	−24	−62	44

Note: The single-step method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE B2 Rhyme data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using Simes family of critical vectors with $\delta \in \{0, 1, 9\}$.

Cluster	Size	% active			Voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$			x	y	z
		$\delta = 0$	$\delta = 1$	$\delta = 9$			
LOC/LG/OFG/PG/SFG/FOC/P/IFG/IC/CG	11,045	67.48%	84.23%	87.27%	4	12	48
Left SPL/PCG	1546	0%	1.55%	2.52%	−24	−62	44

Note: The iterative method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE B3 Rhyme data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using AORC family of critical vectors with $\delta \in \{0, 1, 9\}$.

Cluster	Size	% active			Voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$			x	y	z
		$\delta = 0$	$\delta = 1$	$\delta = 9$			
LOC/LG/OFG/PG/SFG/FOC/P/IFG/IC/CG	3331	68.67%	84.66%	87.65%	4	12	48
Left SPL/PCG	1546	0%	1.68%	2.72%	−24	−62	44

Note: The iterative method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE B4 Rhyme data: Clusters S identified with threshold $t = 3.2$ and $\overline{\pi}(S)$ computed using the family of critical vectors based on the Higher Criticism.

Cluster	Size	% active	Voxel coordinates		
S	$\lvert S\rvert$	$\overline{\pi}(S)$	x	y	z
		Higher Criticism			
LOC/LG/OFG/PG/SFG/FOC/P/IFG/IC/CG	3331	84.1%	4	12	48
Left SPL/PCG	1546	0%	−24	−62	44

Note: The single-step method is applied. The size of the clusters $|S|$ and the voxel coordinates (x, y, z) are reported for each cluster.

TABLE B5 Rhyme data: Clusters S identified by threshold-free cluster enhancement (TFCE) method[20] and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 27$) and parametric ARI.

Cluster	Size	% active		
S	$\lvert S\rvert$	$\overline{\pi}(S)$		
		Perm Simes (B7)	Perm AORC (B8)	Parametric Simes (B9)
LG/OP/P/SFG/FP/I/OFG/LO/PrG/PCG/PG	73,401	44.76%	45.57%	38.78%

Note: The size of the clusters $|S|$ is reported for each cluster.

APPENDIX C. WORD OBJECT ANALYSIS

We analyze the dataset provided by Duncan et al.[51] It consists of 48 subjects looking 4 different visual stimuli: written words, pictures of objects, scrambled pictures of the same objects, and consonant letter strings. It is a block design where each block contains 16 stimuli from a single category using a one-back task, that is, two runs are performed. Therefore, in this case, a third-level analysis was carried out. Table C1 reports the results regarding the contrast of the activation difference between word and consonant string stimuli, considering the single-step method α equals 0.05 and $\delta = 1$. We found activation in Intracalcarine Cortex (IC), Lingual Gyrus (LG), Precentral Gyrus (PrG), Cuneal Cortex (CC), Planum Temporale (PT), Supramarginal Gyrus (SG), Amygdala (A), Superior Temporal Gyrus (STG), Insular Cortex (I), Lateral Occipital Cortex (LO), Middle Frontal Gyrus (MFG), Precuneous Cortex (PrC), Cingulate Gyrus (CG), Accumbens (Ac), Central Opercular Cortex (CO), Thalamus (T), and Superior Frontal Gyrus (SFG).

Figure C1 shows the TDP as a cluster brain map regarding the results using the Simes family confidence bound.

If you are interested in analyzing other possible contrasts, for example, words versus scrambled pictures, you can find the full dataset in https://github.com/angeella/fMRIdata.[40]

TABLE C1 Word-object data: Clusters S identified with threshold $t = 3.2$ and active proportion percentage $\overline{\pi}(S)$ using Simes and AORC families ($\delta = 1$) and parametric ARI, and "drill down" clusters at $t = 4$.

Cluster	Threshold	Size	% active			RFT P-values	Voxel Coordinates		
S	t	$\lvert S\rvert$	$\overline{\pi}(S)$			p_{FWER}	x	y	z
			Perm Simes	Perm AORC	Parametric Simes				
IC/LG/PrC/CC	3.2	17,431	94.52%	94.56%	84.31%	< 0.0001	−2	−78	10
IC/LG/PrC	4	13,469	99.35%	99.35%	97.95%	–	−2	−78	10
Left PT/SG/A/STG/I/LO	3.2	3516	73.23%	73.43%	46.62%	< 0.0001	−24	−14	−12
PT/SG	4	888	90.31%	90.32%	69.93%	–	−56	−44	14
A	4	382	78.27%	78.27%	54.19%	–	−24	−14	−12
STG	4	289	74.74%	74.74%	56.75%	–	−52	−12	−6
I	4	117	56.41%	56.41%	31.62%	–	−34	−16	16
LO	4	44	4.55%	4.55%	0%	–	−42	−64	48
Right MFT/PrG	3.2	2217	74.92%	74.97%	62.65%	< 0.0001	26	6	48
—	4	1658	94.75%	94.75%	83.78%	–	26	6	48
CG	3.2	1640	64.57%	64.57%	52.68%	< 0.0001	−4	−44	32
—	4	1101	92.1%	92.1%	78.47%	–	−4	−44	32
Right STG/A/Ac	3.2	1354	55.1%	55.1%	38.47%	< 0.0001	24	−14	−16
STG	4	345	77.97%	77.97%	54.2%	–	58	−10	−8
A	4	258	66.67%	66.67%	41.47%	–	24	−14	−16
Ac	4	168	66.67%	66.67%	45.23%	–	−4	8	−10
Right CO	3.2	792	29.17%	29.17%	7.2%	< 0.0001	50	−6	8
—	4	270	67.78%	67.78%	21.11%	–	50	−6	8
PG/CG	3.2	637	63.42%	63.42%	53.06%	< 0.0001	4	24	36
—	4	480	84.17%	84.17%	70.42%	–	4	24	36
Right LO	3.2	603	41.79%	41.79%	2.338%	< 0.0001	46	−64	40
—	4	331	73.72%	73.72%	42.6%	–	46	−64	40
Left SFG	3.2	449	41.2%	41.2%	28.06%	< 0.0001	−24	4	46
—	4	266	69.55%	69.55%	47.37%	–	−24	4	46
Right T	3.2	197	17.26%	17.26%	8.63%	0.0003	−20	−26	6
—	4	86	39.53%	39.54%	19.77%	–	−20	−26	6
—	3.2	191	1.57%	1.57%	0%	0.0004	24	−40	20
—	4	47	2.13%	2.13%	0%	–	24	−40	20
Left I	3.2	188	13.82%	13.83%	5.85%	0.0005	−28	16	4
—	4	62	41.94%	41.94%	17.74%	–	−28	16	4
IC	3.2	58	25.86%	25.86%	18.97%	0.084	−32	6	10

Note: The single-step method is applied. The size of the cluster $\lvert S\rvert$ and the voxel coordinates (x, y, z) are reported for each cluster.

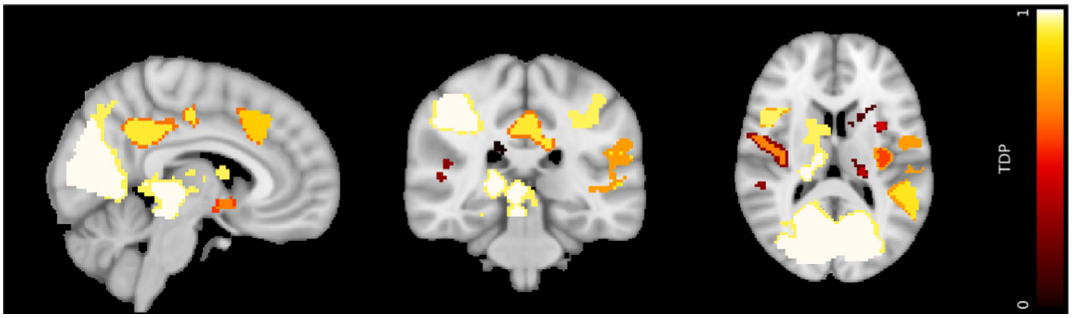

FIGURE C1 Word-object data: True discovery proportion map using the Simes family of critical vectors with $\delta = 1$. Colors express the true discovery proportion for clusters corresponding to a threshold of 3.2 and "drilled" down at 4.

APPENDIX D. ITERATIVE APPROACH

In this section, the iterative version, defined in Theorem 3, is examined following the simulation analysis proposed in Section 9. The method is applied directly on S_{π_1} the set of true discoveries, therefore $\pi(S_{\pi_1}) = 1$.

Figure D1 illustrates the behavior of the approximated iterative method using a different number of combinations. The approximation version becomes exact when the number of combinations goes to infinity. However, as we can see in Figure D1, the results using only 10 combinations are nearly equal to the results using 1000 combinations. In addition, looking at Figure D2, we can see that the method is robust if a different number of variables are considered, that is, the lines in Figure D2 are below 1 (true discovery proportion). In this case, we fix $\theta = 0.2$.

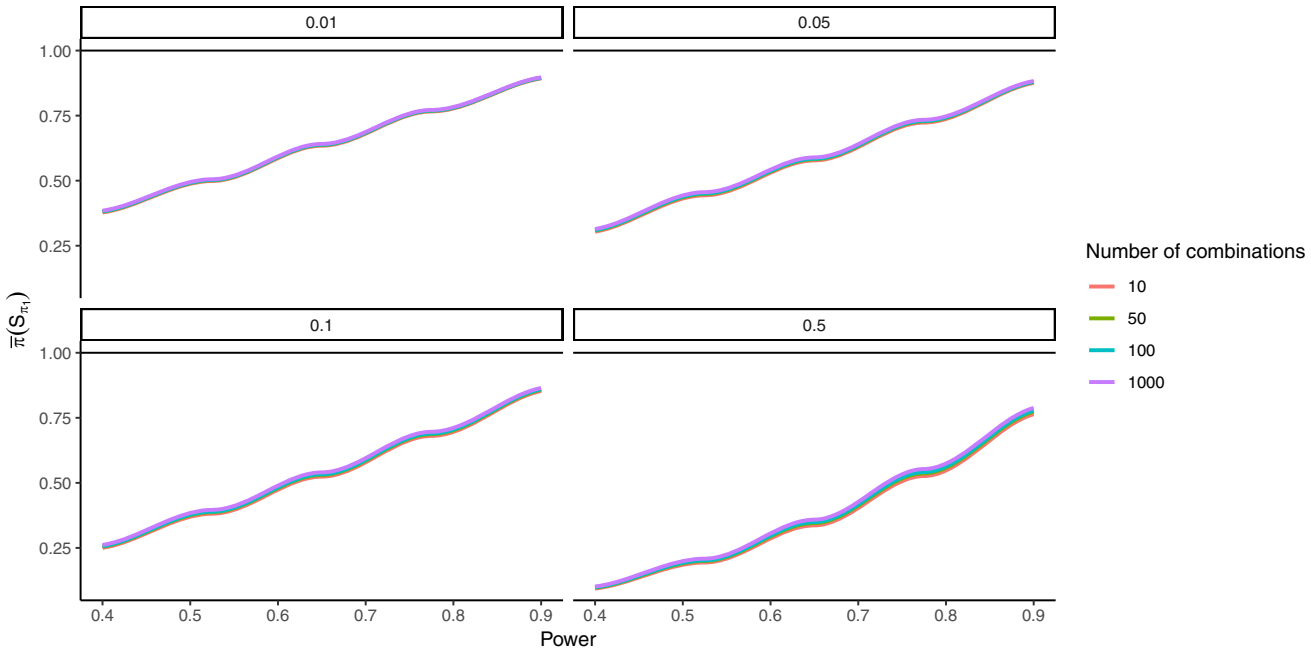

FIGURE D1 Simulated true discovery lower bounds for S_{π_1} over different values of $\theta \in \{0.01,\ 0.05,\ 0.1,\ 0.5\}$ and power using the approximated iterative version with 10, 50, 100, and 1000 random combinations (ie, colored solid lines). The solid black line represents the true discovery proportion $\pi(S_{\pi_1}) = 1$.

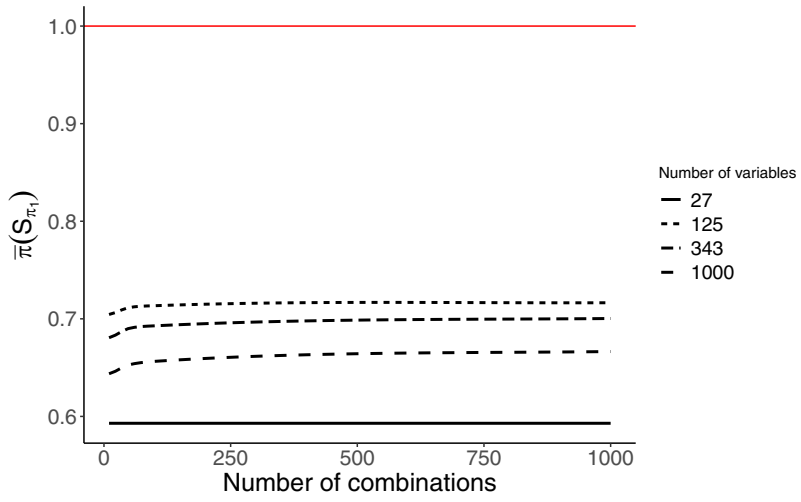

FIGURE D2 Simulated true discovery lower bounds for S_{π_1} over different values of $m \in \{27,125,343,1000\}$ variables using the approximated iterative version with 10, … , 100 random combinations. The solid red line represents the true discovery proportion $\pi(S_{\pi_1}) = 1$.

APPENDIX E. PROOF OF THEOREM 1

Proof. In Lemma 6 by Goeman et al,[18] define $l_{i:n} = l_i$ for every $i \geq 1$ and $n \geq 0$. This lemma then implies that:

$$\max_{1 \leq u \leq |S|} 1 - u + |\{i \in S : p_i \leq l_u\}| \tag{E1}$$

are valid simultaneous bounds, as was to be shown. ∎

APPENDIX F. PROOF OF THEOREM 2

We do not report a formal proof of Theorem 2 since you can directly refer to Hemerik et al [p. 643].[14] Indeed, the relationship between the definition of the λ_α calibration parameter and the power of the method is evident, that is, a large value of λ_α leads to more power. We mainly have rephrased Hemerik et al[14] theorem based on the concept of upper bound for the false discovery proportion in terms of the λ_α calibration parameter. To sum up, Theorem 2 gets an improvement of λ_α, which gives an improved critical vector, and then using Theorem 1 we get an improved lower $(1 - \alpha)$ confidence bound of $a(S)$ simultaneously for all $S \subseteq B$.

APPENDIX G. VALIDATING PERMUTATION-BASED ARI

We propose here the results of performing the one-sample t-tests instead of the two-sample t-tests in the Oulu dataset[3] following the same procedure as Section 8. Figure G1 is structured as Figure 7 presented in Section 8. Again, we can note how the parametric-based ARI returns, in most cases, an estimated FWER equals 0, while the permutation-based ARI gains power controlling the FWER at the same time.

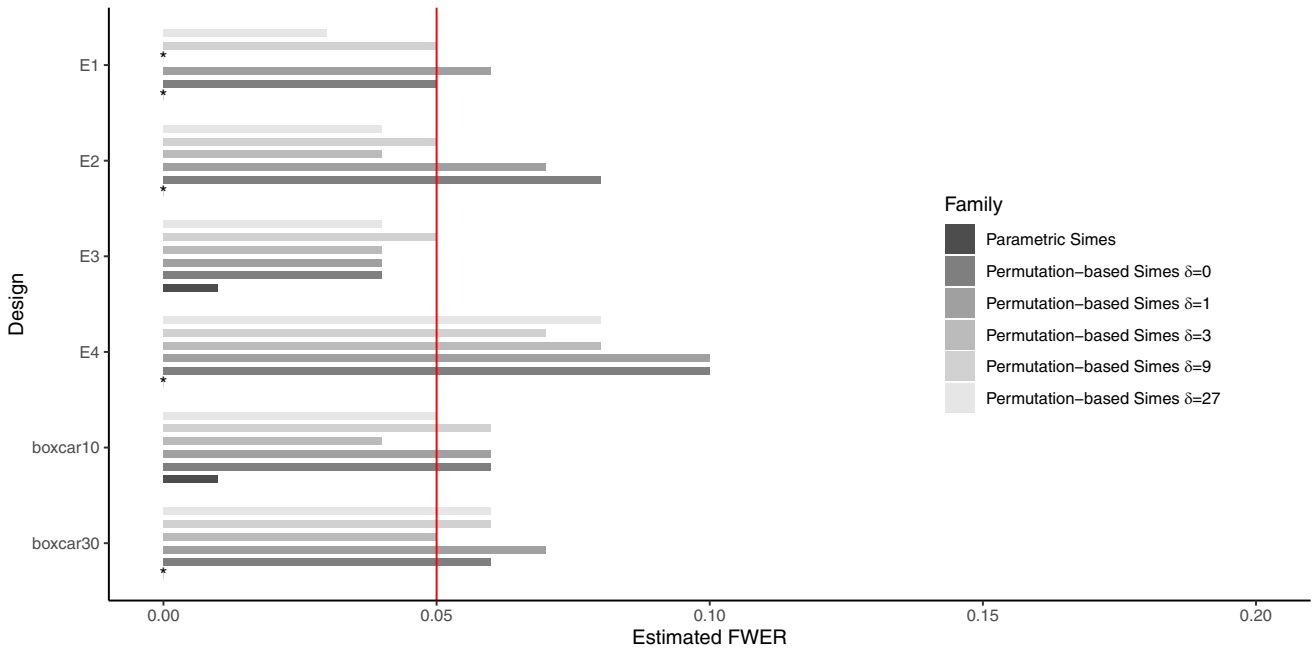

FIGURE G1 Estimated FWER considering six different first level designs, that is, two-block activity paradigms: boxcar10 (10-s on-off), boxcar30 (30-s on-off), and four event activity paradigms, that is, E1 (single event of 2-s activation, 6-s rest), E2 (single event 1- to 4-s activation, 3- to 6-s rest, randomized), E3 (13 events of 3–6 s for each task), and E4 (13 events of 3–6 s for each task, randomized), and six different methods to compute the TDP's lower bound (parametric Simes and permutation-based Simes considering five different values of the shift parameter, that is, $\delta \in \{0, 1, 3, 9, 27\}$). The solid red line represents the estimated nominal FWER equals 0.05, while the star symbols describe estimated FWER equals 0.

APPENDIX H. SIMULATION STUDY

We simulate data considering the simple following model:

$$D_{ij} = \mu_i + \epsilon_{ij}^{\star}$$

where $\mathbf{D}_j \in \mathbb{R}^m$, with $j = 1, \ldots, J$, J is the number of independent observations (ie, subjects) and m is the total number of voxels. The noise $\epsilon_j^{\star} \in \mathbb{R}^m$ follows the multivariate normal distribution with mean 0 and equi-correlation variance structure, that is, $\epsilon_j^{\star} \sim \mathcal{N}(0, \Sigma_{\rho^2})$, where ρ is the level of equi-correlation between pairs of voxels. The signal $\boldsymbol{\mu}$ is computed considering the difference in means having power of the one-sample t-test equals 0.8, that is, $\boldsymbol{\mu} = (z_{1-\alpha/2} + z_{1-\beta})/\sqrt{J}$, where $\alpha = 0.05$ is the significance level, $\beta = 0.8$ is the power level and z_a is the quantiles of the standard normal distribution at level a. The signal $\boldsymbol{\mu}$ is equal to 0 under the null hypothesis.

First of all, we want to understand how the improvement of the nonparametric TDP lower bound changes concerning ρ and the proportion of null hypotheses π_0. Let $J = 50$, $m = 1000$, $\rho^2 \in \{0, 0.01, \ldots, 0.99, 1\}$ and $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$, we simulate data 1000 times and the mean of $\overline{\pi}(S_m)$ over simulation is represented. However, high values of ρ^2 are unrealistic in real applications. For example, the mean correlation across 10,000 randomly sampled voxels equals 0.25 in the case of Rhyme data. The Simes family of confidence bound without shift is taken into account to compare with the parametric approach directly. Having no prior knowledge about the structure of the set of hypotheses to analyze, we consider the full set of hypotheses, that is, S_m. Figure H1 shows the difference of $\overline{\pi}(S_m)$ computed using the permutation and parametric methods over the ρ^2 and π_0 values. As expected, the permutation approach gets some power with respect to the parametric one in the case of correlation between pairs of variables. It can handle any type of dependence structure of the p-values.

Secondly, we want to examine why certain families of critical curves do not provide good results in Section 7. The Higher Criticism critical vector (7), the Beta critical vector (8), and the Simes critical vector (5) are then used to compute $\bar{a}(S_m)$ using simulated data with $\pi_0 = 0.9$, $m = 1000$ and $J = 50$. As previously, we repeat the simulations 1000 times for each framework, and the mean value of $\bar{a}(S_m)$ is computed. Figure H2 shows the behavior of these three families of critical vectors with respect to $\rho^2 \in \{0, 0.01, \ldots, 0.99, 1\}$. In Section 5, we said that the Higher Criticism and Beta families could be problematic in the case of a strong correlation between tests. As expected, the Beta critical vector does not work in the case of a strong correlation between variables. This is also due to computational numerical difficulties, that is when the dashed line in Figure H2 disappears. The Higher Criticism family seems to work, but it loses power with an increase in correlation.

Thirdly, we want to analyze how the Simes family of critical curves (5) works if anti-conservative p-values distribution is considered. Let $J = 50$, $m = 1000$, $\rho = 0$ and $\pi_0 = 0.9$, we compute $\bar{a}(S_m)$ for every 1000 simulations, and once again

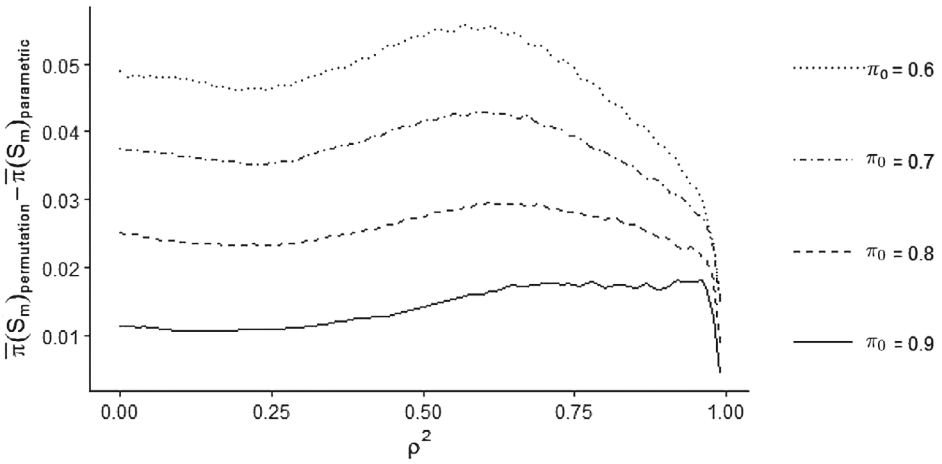

FIGURE H1 Difference of lower bounds for the true discoveries proportion considering the permutation $\overline{\pi}(S_m)_{\text{permutation}}$ and parametric $\overline{\pi}(S_m)_{\text{parametric}}$ methods using simulated data and considering the full set of hypotheses S_m over different values of $\rho^2 \in \{0, 0.01, \ldots, 0.99, 1\}$ and $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$.

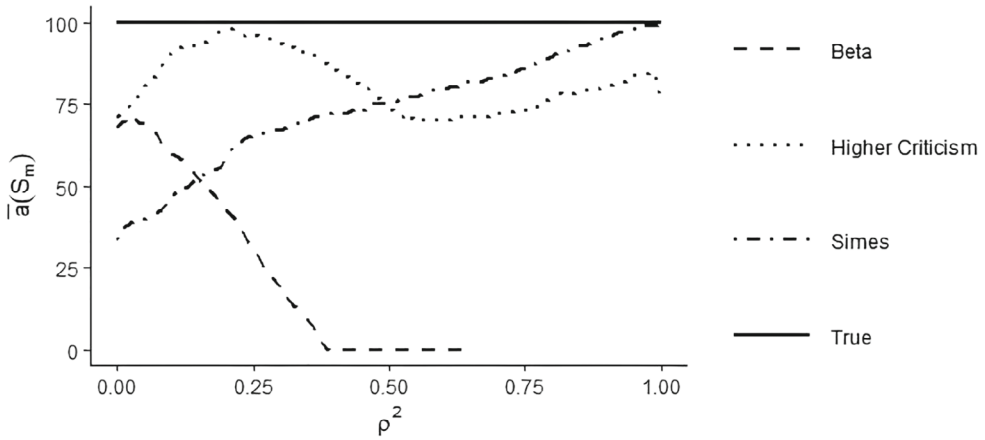

FIGURE H2 Simulated true discovery lower bound over S_m and different values of $\rho^2 \in \{0, 0.01, \ldots, 0.99, 1\}$ using the Higher Criticism (dotted line), Beta (dashed line) and Simes critical vectors (dotted dashed line). The solid line represents the number of true discoveries, which equals 100 considering 1000 variables and the proportion of null hypotheses $\pi_0 = 0.9$.

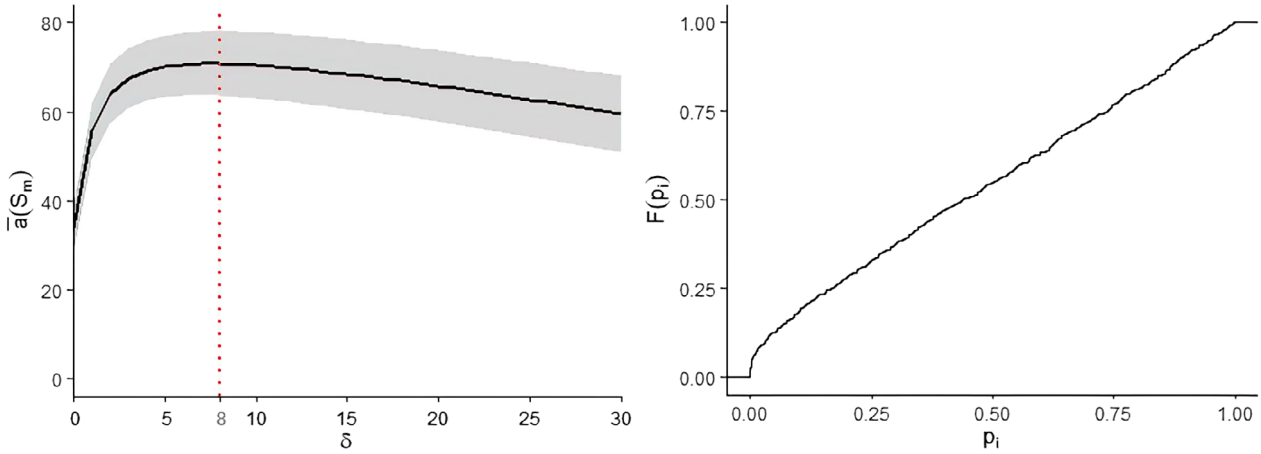

FIGURE H3 Left side: True discovery lower bound using simulated data. The full set of hypotheses S_m is considered over different values of δ. Right side: Empirical cumulative density Function of observed raw p-values, that is, $F(p_i)$.

the mean over simulations is reported. Figure H3 shows $\bar{a}(S_m)$ considering the Simes family using $\delta \in \{0, \ldots, 30\}$. We can note that the shifted version works well in the case of anti-conservative p-values if the corrected value for the tuning parameter δ is chosen, described by the red dotted line, that is, $\delta = 8$.

Therefore, we explore how the Simes family of critical curves (5) works with different values of ρ and S size. The left part of Figure H4 shows the mean of the lower bounds for the TDP considering the full set of hypotheses, that is, S_m, and $\delta \in \{0, 5, 10, 15, 20\}$ over 1000 simulations. We can note that in almost all scenarios, the shifted version outperforms the unshifted ones. The difference gets smaller if ρ^2 increases. However, the situation changes if we compute the TDP for a smaller set of hypotheses than S_m as shown in the right part of Figure H4. In this case, we randomly sample 40 hypotheses from the false null ones, that is, S_{40}.

Then, the performance of the iterative approach proposed in Section 4 is compared with respect to the single-step one presented in Section 3. Let consider directly S_{π_1} the set of true discoveries, therefore $\pi(S_{\pi_1}) = 1$. Figure H5 shows the true discovery proportion $\overline{\pi}(S_{\pi_1})$ computed on 1000 simulated data with π_0 equals 0.9, $\rho^2 \in \{0, 0.2, 0.4\}$ and different levels of power used to simulate the data. In this case, we consider $m = 50$ so that we can use the exact iterative method. First of all, we can see how the approximated iterative version equals the exact one and, more importantly, how both of them uniformly improve the single-step approach.

Figure H6 illustrates the behavior of the approximated iterative method using a different number of combinations. The method is applied directly on S_{π_1} the set of true discoveries, therefore $\pi(S_{\pi_1}) = 1$. The approximation version becomes exact when the number of combinations goes to infinity. However, as we can see in Figure H6, the results using only 10

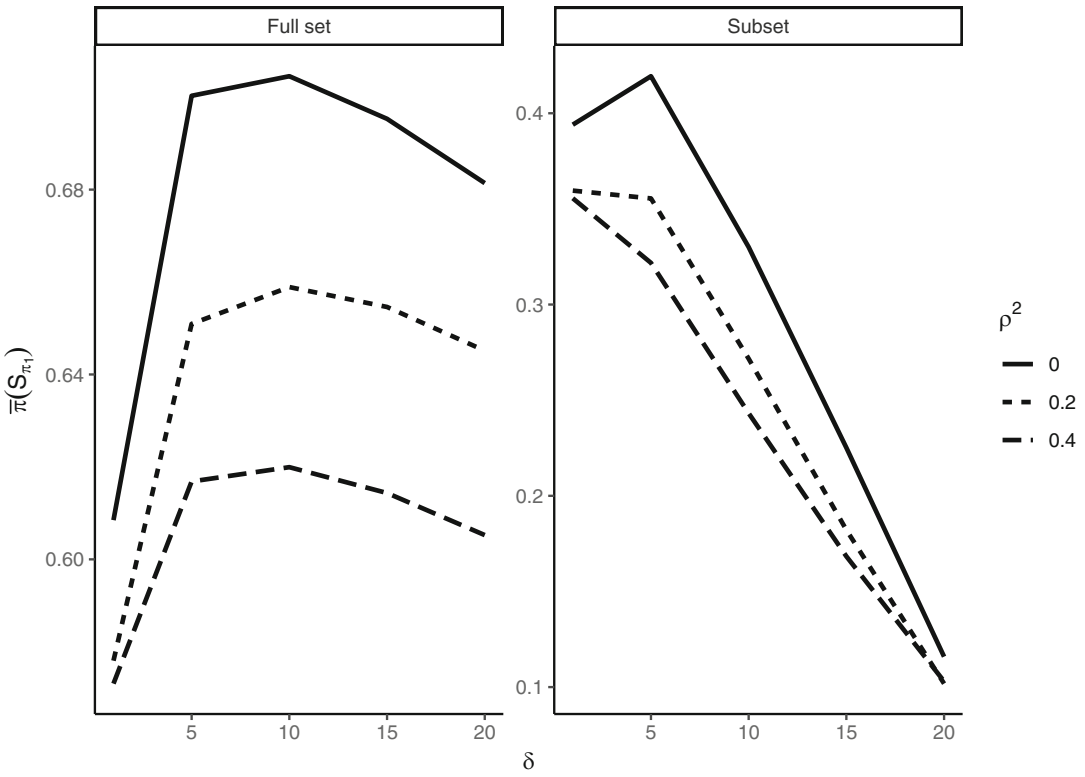

FIGURE H4 Lower Bounds for the true discovery proportion using simulated data the level of equi-correlation $\rho^2 \in \{0, 0.2, 0.4\}$. In the left figure, the full set of hypotheses is considered, while in the right figure a random sample of 40 hypotheses is analyzed. The critical vectors based on the Simes family with $\delta \in \{0, 5, 10, 15, 20\}$ are used in both situations.

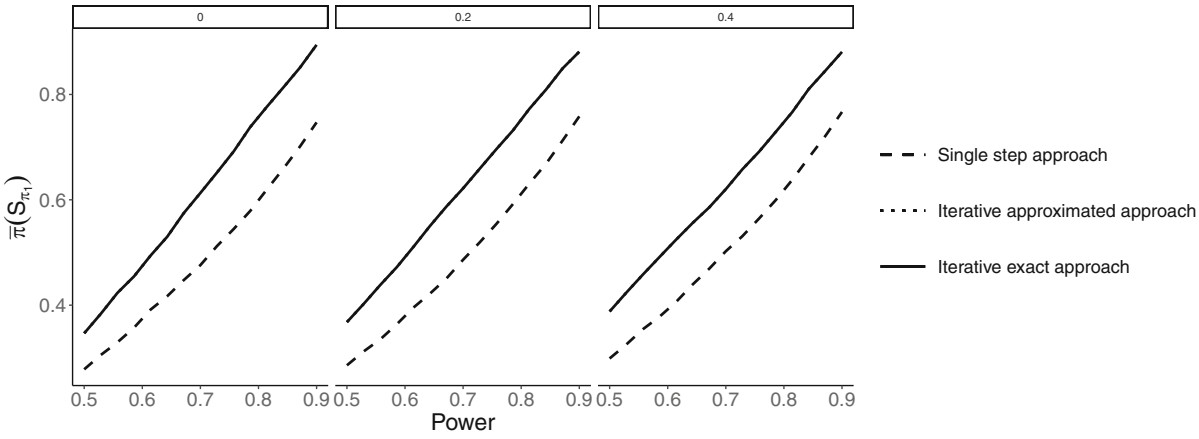

FIGURE H5 Simulated true discovery lower bound for S_{π_1} over different values of ρ^2 and power using the single-step (dashed line), iterative approximated version (dotted line) and iterative exact version (solid line). The dotted line is behind the solid line.

combinations are nearly equal to the results using 1000 combinations. In addition, looking at Figure H7, we can see that the method is robust if a different number of variables are considered, that is, the lines in Figure H7 are below 1 (true discovery proportion).

To sum up, we suggest using the higher criticism and beta families if the correlation across the variables is supposed to be low. Besides, we recommend considering the shifted version of the Simes or AORC family if the interest is in large sets of hypotheses rather than in small ones. and if the distribution of the p-value is expected to be anti-conservative, with a reasonable prior value of δ with respect to the data analyzed.

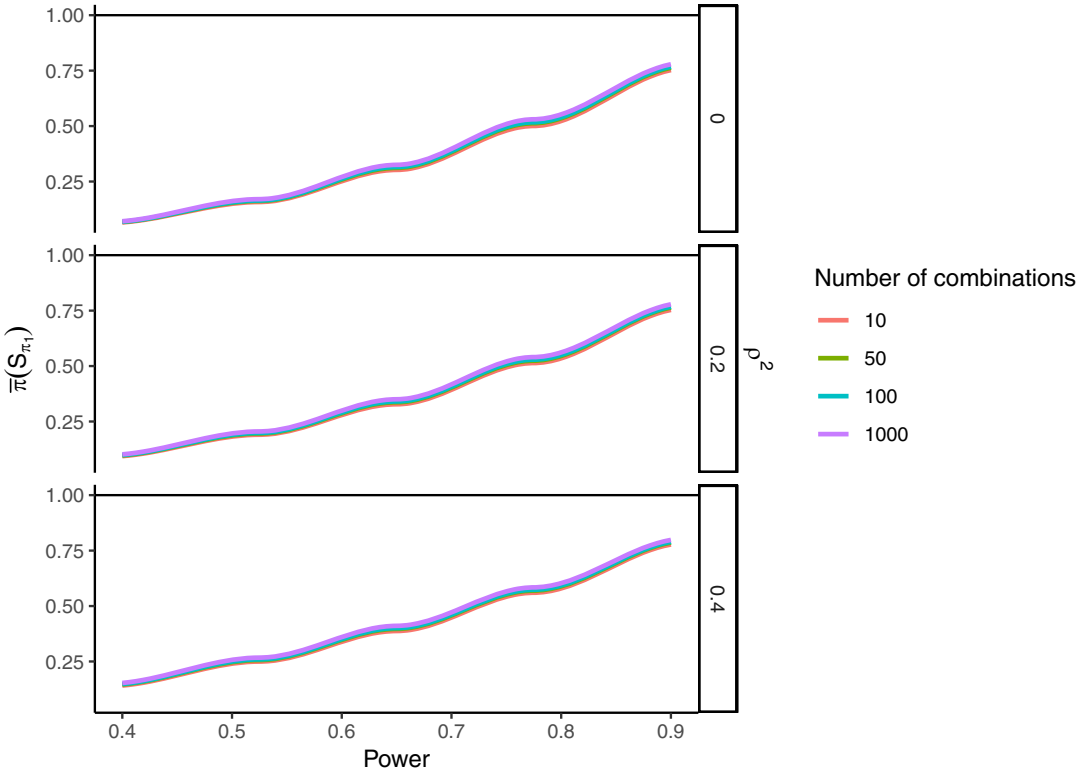

FIGURE H6 Simulated true discovery lower bounds for S_{π_1} over different values of ρ^2 and power using the approximated iterative version with 10, 50, 100, and 1000 random combinations (colored solid lines). The solid black line represents the true discovery proportion $\pi(S_{\pi_1}) = 1$.

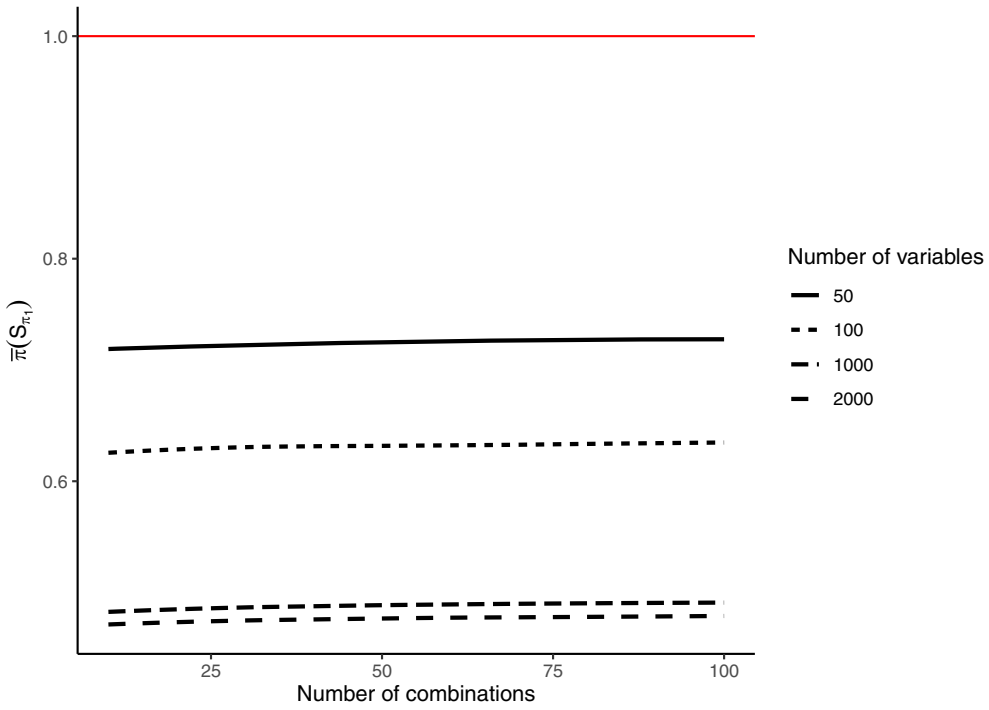

FIGURE H7 Simulated true discovery lower bounds for S_{π_1} over different values of $m \in \{50,100,1000,2000\}$ variables using the approximated iterative version with 10, … , 100 random combinations. The solid red line represents the true discovery proportion $\pi(S_{\pi_1}) = 1$.