Published in final edited form as:
Neuroimage. 2012 December ; 63(4): 1833–1840. doi:10.1016/j.neuroimage.2012.07.040.

Paradoxical results of adaptive false discovery rate procedures in neuroimaging studies

Philip T. Reiss[a,b,*], **Armin Schwartzman**[c,d], **Feihan Lu**[a], **Lei Huang**[e], and **Erika Proal**[f,a]

[a]Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, NY, USA

[b]Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA

[c]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[d]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

[e]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[f]NEUROingenia Clinical and Research Center, Mexico, D.F., Mexico

Abstract

Adaptive false discovery rate (FDR) procedures, which offer greater power than the original FDR procedure of Benjamini and Hochberg, are often applied to statistical maps of the brain. When a large proportion of the null hypotheses are false, as in the case of widespread effects such as cortical thinning throughout much of the brain, adaptive FDR methods can surprisingly reject more null hypotheses than not accounting for multiple testing at all— i.e., using uncorrected *p*-values. A straightforward mathematical argument is presented to explain why this can occur with the *q*-value method of Storey and colleagues, and a simulation study shows that it can also occur, to a lesser extent, with a two-stage FDR procedure due to Benjamini and colleagues. We demonstrate the phenomenon with reference to a published data set documenting cortical thinning in attention deficit/hyperactivity disorder. The paper concludes with recommendations for how to proceed when adaptive FDR results of this kind are encountered in practice.

Keywords

adjusted *p*-values; attention deficit/hyperactivity disorder; cortical thickness; false discovery rate; multiple testing; *q*-values

Introduction

Since the landmark paper of Benjamini and Hochberg (1995) (hereafter, BH), false discovery rate (FDR) methods for inference with large numbers of hypotheses have been

*Corresponding author. Department of Child and Adolescent Psychiatry, New York University School of Medicine, 215 Lexington Ave., 16th floor, New York, NY 10016, USA. Phone: 212-263-3669; fax: 212-263-2476. phil.reiss@nyumc.org.

applied most widely in genome science. The second most frequent domain of application is neuroimaging. The intense interest in FDR-type methods among neuroimagers is attested to by the fact that, on a recent list (Wit, 2010) of the most highly cited papers citing BH, the first paper applying FDR to neuroimaging data (Genovese et al., 2002) ranked fourth.

Here we document and explain a paradoxical phenomenon that can arise in the context of *pointwise adaptive* FDR inference for a statistic image, i.e. a set of test statistics each of which corresponds to a point in the image. The term *pointwise* is meant to encompass statistic images consisting of either voxels, as for PET or fMRI data, or vertices, as for cortical thickness data. *Adaptive* refers to FDR procedures that offer greater power than the original BH procedure. As will be explained below, the BH procedure controls the FDR at a given level times a factor π_0, the proportion of null hypotheses that are true. Since this value is unknown and is near 1 in most practical settings, BH use the approximation $\pi_0 \approx 1$, the effect of which is conservative. Adaptive methods (e.g. Storey, 2002; Benjamini et al., 2006; Liang and Nettleton, 2012) seek an estimate of π_0, which then serves as the basis for a modified, more powerful, FDR procedure.

The paradox is this. Multiple testing adjustment ordinarily means adopting a more stringent threshold to account for the multiple tests performed. But applying adaptive procedures to control the FDR at level α can result in a more lenient threshold, and hence more rejected null hypotheses, than using an uncorrected *p*-value of α to define the threshold. This phenomenon generally occurs when the effect being tested for is widespread throughout the brain, as is common with between-group comparisons of brain structure measures such as cortical thickness (e.g. Narr et al., 2009; Calabrese et al., 2010).

Fig. 1 illustrates how certain distributions of the test statistics can give rise to this phenomenon. The left panel depicts the mixture density with proportion $\pi_0 = 0.9$ of the mass belonging to the $N(0, 1)$ (null) distribution and the rest to the $N(5, 1)$ (non-null) distribution. We simulated 50000 test statistics independently from this mixture—a highly oversimplified model for a "signal" or effect occurring at 10% of the 50000 brain locations for which a statistic is computed, but one that is adequate for this illustration. Thresholding at the 0.05 level by the *q*-value procedure of Storey and Tibshirani (2003), a popular adaptive FDR method, appears to do a good job of separating the null and non-null components. But as π_0 decreases to 0.6 and then to 0.3 (middle and right panels), the *q*-value threshold no longer cleanly divides the distribution into two distinct components. More to the point for our purposes, the value of the threshold decreases with π_0, and for $\pi_0 = 0.3$ it is actually below 1.96, the uncorrected 0.05-level threshold for the null distribution.

Two previous papers have considered adaptive FDR procedures for two different classes of neuroimaging applications, and accordingly have reached quite different conclusions. Logan and Rowe (2004) recommend against adaptive FDR methods in fMRI analysis, on the grounds that π_0 is typically near 1 in that setting and the effect of its estimation on error control is unclear (cf. Example 2 of Schwartzman et al., 2009). On the other hand, in the context of voxel-based morphometry (VBM), π_0 is often well below 1, and hence Chen et al. (2009) see a need for adaptive procedures that increase power by estimating π_0.

The present paper was motivated by a study of grey matter in individuals with attention deficit/hyperactivity disorder, in whomProal et al. (2011) found widespread cortical thinning relative to controls. LikeChen et al. (2009), we found that adaptive FDR procedures have a major impact when studying effects that are spread widely throughout the brain; but the paradoxical results that we obtained lend some support to the cautious stance of Logan and Rowe (2004) toward these procedures.

Background: *p*-values and multiplicity-adjusted *p*-values

To review some concepts that will be important for what follows, we begin with an example. Suppose, for simplicity, that a quantity of interest is taken to be normally distributed with equal variances in each of two populations, such as individuals with and without a disorder, and let the means for the two groups be denoted by μ_1, μ_2. To test the null hypothesis that $H_0 : \mu_1 = \mu_2$ against the alternative $H_A : \mu_1 \neq \mu_2$, we can recruit a sample from each of the groups, measure the quantity for all members of each group, and compute a *t*-statistic. We would ordinarily wish to control the type-I error probability at a prespecified level α, i.e., we choose a *t*-statistic threshold for rejecting H_0 such that, if H_0 is true, the probability that the magnitude of the *t*-statistic exceeds the threshold is α . But rather than simply asking whether or not the observed *t*-statistic exceeds a threshold, we can compute its *p*-value, which can be defined as either (i) the probability, under H_0, of obtaining a test statistic at least as inconsistent with H_0 as the observed one (meaning, in this case, of equal or greater magnitude); or (ii) the smallest significance level α at which H_0 would be rejected (Wright, 1992).

Now consider testing not one but m null hypotheses, e.g., that the two groups have equal means for m different quantities. Classically, the error rate most often controlled is the family-wise error rate (FWER), the probability of falsely rejecting one or more of the null hypotheses. It is easily shown that the Bonferroni procedure, which rejects only those hypotheses whose associated *p*-values lie below α / m, controls the FWER at level α . Equivalently, the Bonferroni procedure can be formulated in terms of adjusted *p*-values (Rosenthal and Rubin, 1983; Wright, 1992), i.e. the lowest FWER level at which the procedure would reject each of the particular null hypotheses. If we let $p_{(1)} \leq \ldots \leq p_{(m)}$ denote the *p*-values sorted in ascending order and $H_{(1)}, \ldots, H_{(m)}$ denote the corresponding null hypotheses, the Bonferroni-adjusted *p*-value for $H_{(i)}$ is $p_{(i)}^{Bon} = m p_{(i)}$.

False discovery rate procedures

Benjamini-Hochberg procedure

Benjamini and Hochberg (1995) argue that in many multiple testing applications, it is more useful to consider not just the probability of any false rejections but the expected number of such rejections. Instead of the FWER, they propose to control the false discovery rate, the expected proportion of errors among the hypotheses rejected by the procedure. Formally, we define a random variable R as the number of hypotheses rejected, and a second random variable V as the number of *true* null hypotheses that are rejected, i.e. the number of false rejections. The FDR is then defined as $E(V/R)$. There is an ambiguity in this definition, however, in that R may equal 0. BH resolve this by setting V/R to 0 when $R = 0$, leading to the more precise definition

$$FDR = Pr(R>0)E(V/R|R>0). \quad (1)$$

To control the FDR at level α, a "step-up" procedure may be used: if $p_{(i)} < i\alpha / m$ for any i, then we reject $H_{(1)}, \ldots, H_{(k)}$ where k is the largest number such that this inequality holds; otherwise none of the hypotheses is rejected. BH show that if the test statistics corresponding to true null hypotheses are independent, this procedure controls the FDR at level $\pi_0 \alpha$, where π_0 is the proportion of null hypotheses that are true. In view of this key result, the step-up procedure, although originally proposed by Simes (1986), is commonly referred to as the "BH procedure" for FDR control.

In practical applications in neuroimaging and other fields, the test statistics are not independent, but Benjamini and Yekutieli (2001) show that the BH procedure controls the FDR at level $\pi_0 \alpha$ under a technical positive dependence condition which is often assumed to hold at least approximately.[1] Since in general $\pi_0 < 1$, the step-up procedure controls the FDR at a level lower than α (under appropriate dependence assumptions). However, since π_0 is unknown and in most applications is close to 1, the BH procedure is generally referred to as controlling FDR at level α, and it remains the standard procedure for this purpose.

Although the BH procedure controls FDR rather than FWER, it can, like the Bonferroni procedure, be usefully framed in terms of adjusted *p*-values (Benjamini et al., 2006). The BH-adjusted *p*-value for $H_{(i)}$, i.e., the smallest FDR level at which $H_{(i)}$ is rejected by the BH procedure, is given by

$$p_{(i)}^{BH} = \min_{i \le j \le m} \frac{m p_{(j)}}{j}. \quad (2)$$

Clearly $p_{(i)}^{BH} \le p_{(i)}^{Bon}$, and hence the BH procedure will always reject at least as many hypotheses at a given (FDR) level α as the Bonferroni procedure will reject at (FWER) level α .

An adaptive two-stage FDR procedure

As noted in the Introduction, many authors have sought to develop adaptive FDR procedures. These procedures seek to improve power by estimation of the true null proportion π_0 (an idea that predates FDR; see Schweder and Spjøtvoll, 1982).

For example,Benjamini et al. (2006) propose the following two-stage step-up procedure:

1. Apply the BH step-up procedure at level $\frac{\alpha}{1+\alpha}$.
2. If either all or none of the m hypotheses are rejected in step 1, this is the final conclusion, and the procedure terminates. Otherwise, estimate π_0 by $\hat{\pi}_0 = 1 - r/m$, where r is the number of hypotheses rejected in step 1.
3. Apply the BH step-up procedure at level $\frac{\alpha}{\widehat{\pi}_0(1+\alpha)}$

Benjamini et al. (2006) prove that, if the test statistics are independent, this two-stage procedure controls the FDR at level α . They also provide empirical evidence that this control is maintained under positive dependence.

Positive FDR and q-values

Storey and colleagues (Storey, 2002, 2003; Storey and Tibshirani, 2003; Storey et al., 2004) introduce a modified approach to FDR that differs from the original BH procedure in two key respects: an estimate of π_0 is employed, and the error rate of interest is $E(V/R|R > 0)$, the *positive FDR* (pFDR). Storey (2002) argues that pFDR is a more appropriate error rate to consider than the original FDR (1).

Any given rejection region (e.g., reject hypotheses for which the *t*-statistic has magnitude at least 3, or the raw *p*-value is at most 0.001) has an associated pFDR. Indeed, in some reports (e.g., Chiang et al., 2007; Lepore et al., 2008), a fixed threshold such as 0.01 is chosen for

[1] They also show that, irrespective of the dependence structure, the procedure always controls the FDR at level $\pi_0 \alpha \sum_{i=1}^{m} 1/i$.

the raw p-values, and the pFDR associated with that threshold is reported. But Storey (2002) argues that it is more appropriate to let the observed p-values determine the rejection region. To that end he proposes a new quantity, the q-value, as the basis for thresholding. The standard way to apply Storey's paradigm, and the one advocated in the neuroimaging context byChen et al. (2009), is to reject those hypotheses with q-values below a specified level such as 0.05.

The q-value of an observed statistic t is the minimum pFDR arising from a rejection region that contains t. Under certain assumptions, most notably independence of the test statistics, the q-value can also be understood as the posterior probability that a null hypothesis is true, given a test statistic at least as extreme as the one observed (the reverse of definition (i) above of a p-value; Storey, 2003).[2] A key special case is when we take the observed raw p-values themselves as the test statistics: Storey (2002) then defines the q-value associated with $p_{(i)}$ ($i = 1, \ldots, m$)—again, under the assumption that these p-values are independent[3]—as the smallest pFDR associated with some threshold $\gamma \geq p_{(i)}$. With some additional technical assumptions (Storey, 2002), the q-value corresponding to $p_{(i)}$ can be expressed as

$$q(i) = q(p_{(i)}) = \min_{\gamma \geq p_{(i)}} \frac{\pi_0 \gamma}{\Pr(P \leq \gamma)}, \quad (3)$$

where the denominator refers to the probability that the p-value for a given test is at most γ.

Note that (3) is an unobserved quantity; the "q-values" $\hat{q}_{(1)} \leq \ldots \leq \hat{q}_{(m)}$ reported in practice are more precisely q-value *estimates*. The main step in estimating (3) is to obtain a null proportion estimate $\hat{\pi}_0$; see Appendix A. Upon completing this step, the estimated q-value for $H_{(i)}$ is

$$\widehat{q}_{(i)} = \min_{i \leq j \leq m} \frac{m \widehat{\pi}_0 p_{(j)}}{j}. \quad (4)$$

Asymptotic conservatism of this q-value estimate is demonstrated byStorey et al. (2004).[4] It is worth noting that $\hat{q}_{(i)}$ would reduce to the BH FDR-adjusted p-value (2) if we took $\hat{\pi}_0 = 1$; or from an algorithmic standpoint, a q-value threshold of α can be applied by performing the BH procedure at level $\alpha / \hat{\pi}_0$.

Paradoxical behavior of adaptive FDR procedures

We can now explain mathematically the anomalous behavior sometimes exhibited by adaptive FDR procedures. By (4), if $\hat{\pi}_0 = 1$ then the q-value is always larger than the corresponding unadjusted p-value. But if $\hat{\pi}_0 < 1$, then for all $i > m\hat{\pi}_0$,

$$\widehat{q}_{(i)} \leq m \widehat{\pi}_0 p_{(i)} / i < p_{(i)}, \quad (5)$$

i.e., the ith-smallest q-value is smaller than the corresponding p-value. When $\hat{\pi}_0 \approx 1$, $i > m\hat{\pi}_0$ will occur only for i near m, so inequality (5) is entailed only for the largest p-values,

[2] In this Bayesian interpretation, π_0 is the prior probability of the null hypothesis. A small value of π_0 then implies that the prior information no longer supports the null for most tests; hence the tendency to reject a large number of null hypotheses.

[3] Strictly speaking, the q-value is undefined if the test statistics are not independent, but in practice this requirement is glossed over. Some empirical evidence (e.g., Storey, 2003; Storey et al., 2004) supports the use of q-values for dependent data; cf. the brief remarks above concerning dependence and the BH procedure.

[4] Somewhat confusingly, whereas the q-value is defined in terms of pFDR, the fractional expression in the q-value estimate (4) is an estimate of FDR. This follows Storey and Tibshirani (2003), whereas the q-value estimate in Storey (2002) is based on an estimate of pFDR. SeeStorey et al. (2004), p. 196, regarding these alternative estimators of the q-value.

which will generally be far from the chosen significance threshold. If, for example, $\hat{q}_{(i)} = 0.93 < 0.95 = p_{(i)}$ for some i near m, this may be somewhat counterintuitive, but it has no effect on declarations of significance since neither the raw p-value nor the q-value points toward rejecting $H_{(i)}$.

As $\hat{\pi}_0$ decreases, however, $i > m\hat{\pi}_0$ for progressively smaller values of i, and hence the p-values for which (5) must hold move toward the left tail of the distribution. Thus for very small $\hat{\pi}_0$ it is entirely possible to have i for which, say, $p_{(i)} = .07$ but $\hat{q}_{(i)} = .03$. Adopting the conventional 0.05 significance level, multiple testing correction would then have the highly counterintuitive effect of converting the evidence against $H_{(i)}$ from "non-significant" to "significant"!

This surprising result cannot occur with the BH procedure. A BH-adjusted p-value can never be less than the unadjusted p-value since, by (2), $p^{BH}_{(i)} \geq \min_{i \leq j \leq m} p_{(j)} = p_{(i)}$. For the two-stage procedure ofBenjamini et al. (2006), the phenomenon is more difficult to demonstrate mathematically, since there is no explicit expression for an adjusted p-value analogous to (2) or (4). But the simulation results presented next show that this paradoxical behavior can indeed occur with the two-stage procedure.

An illustrative simulation study

Setup

We conducted a simulation study to examine the performance of the three FDR approaches described above for a large number of two-sample t-tests, when for a large proportion of hypotheses the two groups' distributions differ by a location shift Δ.

We fixed the sample size at 25 per group and the number of variables at 1000. For members of group 1, each variable's values were simulated from the $N(0, 1)$ distribution. For group 2, values were simulated from either the $N(0, 1)$ distribution for $1000\pi_0$ variables, and from the $N(\Delta, 1)$ distribution, for a positive Δ, for the remaining $1000(1-\pi_0)$ variables. The simulation settings were arrayed in a 2 × 2 grid:

- In the first two sets (see panels (a) and (b) of Fig. 2), Δ was fixed at 0.7, and 299 replications were performed with each of the values $\pi_0 = 0, 0.1, 0.2, \ldots, 1$.
- In the last two sets (see panels (c) and (d) of Fig. 2), π_0 was fixed at 0, and 299 replications were performed with each of the values $\Delta = 0, 0.1, 0.2, \ldots, 1$.
- In the first and third sets, the 1000 variables were mutually independent.
- In the second and fourth sets, the root mean square (RMS) of the between-variable correlations was approximately 0.15. The data were generated using R code (R Development Core Team, 2011) accompanying Efron (2010) (available at http://stat.stanford.-edu/~omkar/monograph/data.html), which attains a desired RMS correlation by simulating values that are highly correlated within each of a number of equal-sized blocks (the default, which we used, is 5 blocks).

For each replication, we determined the proportion of null hypotheses rejected at the (two-sided) $\alpha = 0.05$ level, out of the 1000 t-tests performed, by each of the following methods:

1. uncorrected two-sided p-value;
2. the original step-up procedure of Benjamini and Hochberg (1995);
3. the two-stage step-up procedure ofBenjamini et al. (2006);
4. the q-value procedure of Storey and Tibshirani (2003).

The simulations were performed in R, with the `qvalue` package (Dabney et al., 2012) used for the *q*-value method.

Results

As can be seen by comparing the two left subfigures of Fig. 2 with the two right subfigures, correlation among the test statistics increases the variance of the number of rejections, but otherwise does not affect the pattern of results.

The two upper subfigures display results with the effect size Δ fixed at 0.7 and the proportion of false null hypotheses ($1-\pi_0$) increasing from 0 to 1. As one would expect, the mean proportion of rejections based on uncorrected *p*-values increases linearly with $1 - \pi_0$. Consistent with the discussion above, the BH procedure never rejects more hypotheses than uncorrected *p*-values do, but the *q*-value method curve begins to exceed the *p*-value curve around $1-\pi_0 = 0.6$. Consistent with the findings of Chen et al. (2009), the two-stage FDR method lies in between the BH and *q*-value methods, and its curve crosses the raw *p*-value curve only for the highest values of $1 - \pi_0$.

Qualitatively similar results are seen in the two lower subfigures of Fig. 2 for the simulations with $\pi_0 = 0$ and varying Δ. We showed above that adaptive FDR can reject more null hypotheses than unadjusted *p*-values when an effect is widespread, i.e., π_0 is small. The lower subfigures illustrate how similar results can also occur when $\pi_0 = 0$—meaning that the null/non-null mixture is no longer a correct description of the data—if the effect is strong enough. We shall return to this point in the Discussion.

Longitudinal ADHD study

Background

Proal et al. (2011) studied cortical thickness and voxel-based morphometry in 59 adults in whom attention deficit/hyperactivity disorder (ADHD) had been established in childhood, along with 80 controls. Both groups consisted of Caucasian males. The ADHD group had been recruited 33 years earlier, while the comparison group had been recruited ten years thereafter, but was group-matched for race, childhood socioeconomic status, and place of residence. The MRI scans were performed at the most recent of three clinical follow-up visits (mean ages 18.4, 25.0 and 41.2). Despite the long elapsed time since enrollment, the authors found widespread grey matter deficits in the ADHD group relative to the controls.

Here we apply the three FDR methods discussed above to comparison of cortical thickness in the two groups at each of 81924 vertices. The test statistic at each vertex is a Wald *t*-statistic for the group effect on cortical thickness, controlling for two covariates: age at time of the scan, and scanner.[5] Negative *t*-statistics imply lower cortical thickness in ADHD subjects than in controls. Upon converting these *t*-statistics to standard normal statistics, the distribution of values across the brain is summarized by the histogram in Fig. 3 (see below for further discussion of this figure).

FDR results

The left subfigure of Fig. 4 displays the spline smoothing method by which π_0 is estimated (see Appendix A). Extrapolating the curve to $\lambda = 1$ yields the final estimate $\hat{\pi}_0 = 0.234$. Plugging this estimate into (4) yields the *q*-values, which are plotted against the raw two-

[5]Two different scanners were used in the study. Aside from controlling for scanner, we applied a chi-square test to the cross-tabulation of diagnostic group by scanner. The nonsignificant result ($\chi^2_1=0.62$ $p = 0.43$) diminishes the concern that observed group effects may be attributable to confounding between scanner and diagnosis.

sided *p*-values in the right subfigure, revealing an instance of the paradox we have described: *p*-values up to 0.128 ($|t| > 1.53$) correspond to *q*-values below 0.05, and hence the null hypothesis is rejected at the 0.05 level for 49161 of the 81924 vertices by the *q*-value method, versus 36569 rejections based on uncorrected *p*-values. The other two methods exhibit conventional behavior for multiple testing procedures, rejecting fewer hypotheses at the 0.05 level than do raw *p*-values: 18004 for BH, and 21772 by the two-stage FDR method.

At most vertices for which the null hypotheses were rejected, the ADHD group exhibited lower cortical thickness, but higher thickness in the ADHD group was found at 2 vertices by the BH procedure, 126 by the two-stage procedure, and 1136 by the *q*-value procedure. Note, however, that the *q*-value method finds so many "significantly positive" vertices only because the preponderance of *negative t*-statistics results in an unusually low threshold (1.53, as mentioned), which implies that *positive t*-statistics above 1.53 are deemed significant. Maps displaying the regions of lower cortical thickness in the ADHD group are presented in Fig. 5.

Somewhat disquieted by the fact that the *q*-value method rejected more null hypotheses than the uncorrected *p*-values, we considered three alternative approaches to analyzing this data set.

Alternative approach 1: Empirical null methodology

Efron (2004, 2010) has developed an empirical Bayes approach to FDR inference in which, as in Storey and colleagues' approach, one aims to estimate each test statistic's posterior probability of having arisen from the null hypothesis distribution. Efron's key insight is that the portion of the observed test statistic distribution corresponding to true null hypotheses is often inconsistent with the theoretical null distribution. Consequently, more accurate inference can be achieved by modeling the test statistic distribution as a mixture of an "empirical null" component, which may differ from the theoretical null, and an "alternative" component. A crucial assumption of the empirical null approach is that $\pi_0 \approx 1$, and the violation of that assumption limits the approach's utility for our data set, as we now show.

Fig. 3 displays a null/non-null mixture distribution for our data, as estimated by the method of Muralidharan (2009). The specifics of this author's empirical null method lie beyond the scope of this paper, but in essence, the portions of the distribution deemed most inconsistent with the null hypothesis are those in which the alternative density, shown in green, accounts for most of the total density. In this instance, Fig. 3 implies that the most significant (or "interesting," in Efron's terminology) test statistics are the small positive values in the right tail. While it is true that values in this range stand apart somewhat from the bulk of the distribution, it would be very strange to report as significant/interesting those regions with the *smallest* between-group differences.

The empirical null approach was developed for large-scale multiple testing settings in which the null portion of the test statistic distribution deviates from the theoretical null for reasons such as unobserved covariates and correlation among the test statistics. The utility of this methodology for neuroimaging applications in which $\pi_0 \approx 1$ has been demonstrated bySchwartzman et al. (2009). But the empirical null approach is not designed for effects that occur across wide portions of a collection of tests, and in our instance, it simply does not provide a useful result.

Alternative approach 2: Adjusting for mean thickness

In neuroanatomic studies in which global effects are observed, it is sometimes appropriate to include the total or mean value across the brain as a covariate in the voxel- or vertex-wise

regressions.Peelle et al. (2012) refer to this approach as "local covariation," and point out that it changes the scientific question being posed. In our context, controlling for whole-brain mean cortical thickness means that we are no longer simply asking, for each vertex, whether a group effect on cortical thickness is observed. Instead we are asking whether there is a group effect beyond that which can be explained by the association between thickness at the given vertex and thickness for the brain as a whole.

For our data, with mean thickness included as a covariate, none of our three FDR procedures declared any vertices significant at the 0.05 level. Thus, like the more sophisticated empirical null method, the covarying approach does not pinpoint brain regions exhibiting a particularly strong group effect. The problem with reporting this as our primary analysis is that we would be missing the main story: widespread cortical thinning in ADHD.

Alternative approach 3: Comparing mean thickness

The simplest alternative is in a sense the opposite of the previous approach: to eschew multiple testing altogether and compare cortical thickness averaged over the entire brain for the 80 controls versus the 59 ADHD individuals. When we regress mean cortical thickness on group, controlling for age and scanner as in the vertexwise analyses, belonging to the ADHD group is found to predict a deficit of 0.031 mm ($t_{135} = -3.24$, $p = .0015$); see also the box plots in Fig. 6 (which do not control for the above two covariates).

The problem with simply comparing the two groups' whole-brain mean thickness is that it does not give rise to a thresholded map identifying the particularly salient regions. On the other hand, finding a significant difference in overall cortical thickness provides a justification for reporting at least an unthresholded map. And if the alternatives for thresholding are overconservative BH results or paradoxical adaptive FDR results, then an unthresholded map may be the most reasonable option.

Discussion

What should be made of an analysis in which adaptive FDR procedures lead to more rejected null hypotheses than do unadjusted *p*-values? It depends, perhaps, on how seriously one takes the "two-groups model" (Efron, 2008), which partitions the test statistic distribution into null and non-null components. In this framework, as noted above, the pFDR is the posterior probability of belonging to the null rather than the non-null component. But in the simulations whose results appear in the lower half of Fig. 2, there is no null component. Accordingly, there are no false discoveries by definition, and in that sense the number of rejections by the *q*-value method is not too high; indeed, no number of rejections would be too high.

On the other hand, perhaps small estimates of π_0 should prompt us to question the applicability of the two-groups model. As Efron (2008) notes in reference to microarray analyses, "Scientific context, which says that there is likely to be a large group of (nearly) unaffected genes… is what makes the two-groups model a reasonable Bayes prior." If, on the other hand, the data contradict the assumption that most of the test statistics are generated by the theoretical null distribution—as evidenced by a small estimate of π_0—there may be little scientific reason to posit a null/non-null mixture model at all. And the less we trust this model, the less reason we have for focusing on the probability of belonging to the null component, i.e., on the pFDR.

The sometimes-paradoxical behavior of adaptive FDR procedures is noted by Benjamini and Hochberg (2000), who present a toy example. But the phenomenon remains little known,

evidently because in most applications $\pi_0 \approx 1$. Benjamini and Hochberg (2000) write that rejecting a hypothesis whose raw p-value exceeds the specified FDR level

> need not be of concern, since we view the various comparisons in a simultaneous framework, and against the background of the many hypotheses rejected in such a study controlling the FDR allows to reject such an hypothesis as well.

Still, they advise that one may wish to avoid such an outcome by imposing a maximum p-value (which may be the chosen FDR level) required for rejecting a hypothesis.

In a study of cortical thinning in ADHD, Narr et al. (2009) report that "the estimated FDR for regions with uncorrected $p < .05$ is 0.0006." This appears to be an instance of what we have called a paradoxical result: i.e., a raw p-value of 0.05 corresponds to an adaptive FDR estimate of 0.0006. Thus the authors, in reporting results meeting the $p = 0.05$ threshold, have implicitly adopted the above suggestion of Benjamini and Hochberg (2000). But in such instances, readers may wonder (as we initially did) why the effect of multiple testing correction is to increase, rather than decrease, significance.

We believe that investigators can forestall such confusion by looking carefully at the results —for example, by means of a histogram of test statistics as in Fig. 3—and considering the assumptions made by different multiple testing methods: e.g., Efron's two-component mixture model is likely inappropriate when π_0 is small. More specific recommendations will naturally depend on the particular application, but the following points seem worth considering:

1. Using the BH procedure instead of adaptive procedures, as inProal et al. (2011), avoids paradoxical outcomes, but may be overconservative.
2. The two-stage FDR procedure ofBenjamini et al. (2006) may be an suitable compromise between the BH and q-value procedures.
3. Comparing whole-brain patterns (alternative approach 3 above) replaces the large-scale multiple testing problem with a single test, likely with very high power. This makes it difficult to present a thresholded map, but may serve as justification for presenting an unthresholded map.
4. If paradoxical adaptive FDR results are reported, this should be clearly noted and explained.

It is our hope that this note will contribute to clear understanding of adaptive FDR procedures, and to thoughtful application of these powerful tools, by analysts of neuroimaging data.

Acknowledgments

We thank Xavier Castellanos, Eva Petkova, Catherine Sugar, Martin Lindquist and Jason Lerch for very helpful discussions; and Yin-Hsiu Chen for bringing a number of key references to our attention; and two anonymous referees, for suggesting a number of improvements in the manuscript. Philip Reiss's research was supported in part by National Science Foundation grant DMS-0907017 and National Institutes of Health (NIH) grant R01 EB009744-01A. Armin Schwartzman's research was partially supported by NIH grants 1R21 EB012177-01A1 and 1P01 CA134294-01. The cortical thickness study was funded by NIH grant RO1 DA016979.

Appendix A. Null proportion estimation

For estimating π_0, Storey and Tibshirani (2003) propose a smoothing method, whereas Storey et al. (2004) offer a bootstrap method. Here we summarize the former, which is the default in the package `qvalue` (Dabney et al., 2012) for R. The p-values for the $m\pi_0$ true null hypotheses are expected to be uniformly distributed between 0 and 1. Hence for any λ

$\in (0, 1)$, approximately $m\pi_0(1-\lambda)$ true null hypotheses should have p-values above $1-\lambda$. On the other hand, for sufficiently large λ, essentially none of the p-values for false null hypotheses should be above $1-\lambda$. This motivates the null proportion estimator

$$\widehat{\pi}_0(\lambda)=\frac{\neq\{i\in\{1,\ldots,m\}:p_i>\lambda\}}{m(1-\lambda)}. \quad \text{(A.1)}$$

Arguing that $\hat{\pi}_0(\lambda)$ should have decreasing bias but increasing variance as λ approaches 1, Storey and Tibshirani (2003) propose to compute $\hat{\pi}_0(\lambda)$ for a range of λ values from 0 to near 1; apply natural cubic spline smoothing with 3 degrees of freedom to these values, to obtain a smooth function $\hat{f}(\lambda)$; and estimate π_0 by the extrapolated value $\hat{\pi}_0 = \hat{f}(1)$.

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B. 1995; 57:289–300.

Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics. 2000; 25:60–83.

Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 2006; 93:491–507.

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

Calabrese M, Rinaldi F, Mattisi I, Grossi P, Favaretto A, Atzori M, Bernardi V, Barachino L, Romualdi C, Rinaldi L, Perini P, Gallo P. Widespread cortical thinning characterizes patients with MS with mild cognitive impairment. Neurology. 2010; 74:321–328. [PubMed: 20101038]

Chen S, Wang C, Eberly LE, Caffo BS, Schwartz BS. Adaptive control of the false discovery rate in voxel-based morphometry. Human Brain Mapping. 2009; 30:2304–2311. [PubMed: 19034901]

Chiang MC, Reiss AL, Lee AD, Bellugi U, Galaburda AM, Korenberg JR, Mills DL, Toga AW, Thompson PM. 3d pattern of brain abnormalities in williams syndrome visualized using tensor-based morphometry. NeuroImage. 2007; 36:1096–1109. [PubMed: 17512756]

Dabney A, Storey JD. q-value: *q*-value estimation for false discovery rate control. R package version 1.28.0. 2012 with assistance from G. R. Warnes.

Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association. 2004; 99:96–104.

Efron B. Microarrays, empirical bayes and the two-groups model. Statistical Science. 2008; 23:1–22.

Efron, B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. New York: Cambridge University Press; 2010.

Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage. 2002; 15:870–878. [PubMed: 11906227]

Lepore N, Brun C, Chou YY, Chiang MC, Dutton RA, Hayashi KM, Luders E, Lopez OL, Aizenstein HJ, Toga AW, Becker JT, Thompson PM. Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors. IEEE Transactions on Medical Imaging. 2008; 27:129–141. [PubMed: 18270068]

Liang K, Nettleton D. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. Journal of the Royal Statistical Society: Series B. 2012; 74:163–182.

Logan BR, Rowe DB. An evaluation of thresholding techniques in fMRI analysis. NeuroImage. 2004; 22:95–108. [PubMed: 15110000]

Muralidharan O. mixfdr: Computes false discovery rates and effect sizes using normal mixtures. R package version 1.0. 2009

Muralidharan O. An empirical Bayes mixture method for effect size and false discovery rate estimation. Annals of Applied Statistics. 2010; 4:422–438.

Narr KL, Woods RP, Lin J, Kim J, Phillips OR, Del'Homme M, Caplan R, Toga AW, McCracken JT, Levitt JG. Widespread cortical thinning is a robust anatomical marker for attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry. 2009; 48:1014–1022. [PubMed: 19730275]

Peelle JE, Cusack R, Henson RNA. Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging. NeuroImage. 2012; 60:1503–1516. [PubMed: 22261375]

Proal E, Reiss PT, Klein RG, Mannuzza S, Gotimer K, Ramos-Olazagasti MA, Lerch JP, He Y, Zijdenbos A, Kelly C, Milham MP, Castellanos FX. Brain gray matter deficits at 33-year follow-up in adults with attention-deficit/hyperactivity disorder established in childhood. Archives of General Psychiatry. 2011; 68:1122. [PubMed: 22065528]

R Development Core Team. R Foundation for Statistical Computing. Vienna, Austria: 2011. R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0

Rosenthal R, Rubin DB. Ensemble-adjusted *p* values. Psychological Bulletin. 1983; 94:540–541.

Schwartzman A, Dougherty RF, Lee J, Ghahremani D, Taylor JE. Empirical null and false discovery rate analysis in neuroimaging. NeuroImage. 2009; 44:71–82. [PubMed: 18547821]

Schweder T, Spjøtvoll E. Plots of *p*-values to evaluate many tests simultaneously. Biometrika. 1982; 69:493–502.

Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73:751–754.

Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B. 2002; 64:479–498.

Storey JD. The positive false discovery rate: a Bayesian interpretation and the *q*-value. Annals of Statistics. 2003; 31:2013–2035.

Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society: Series B. 2004; 66:187–205.

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003; 100:9440–9445.

Wit E. Comment on Y. Benjamini, "Discovering the false discovery rate". Journal of the Royal Statistical Society: Series B. 2010; 72:410–412.

Wright SP. Adjusted *p*-values for simultaneous inference. Biometrics. 1992; 48:1005–1013.

Highlights

- Adaptive false discovery rate procedures boost power via a null proportion estimate.
- But for neuroimaging data, they can reject more null hypotheses than ignoring multiplicity.
- We explain this phenomenon and illustrate it with a cortical thickness example.
- We propose strategies for handling such paradoxical results in practice.

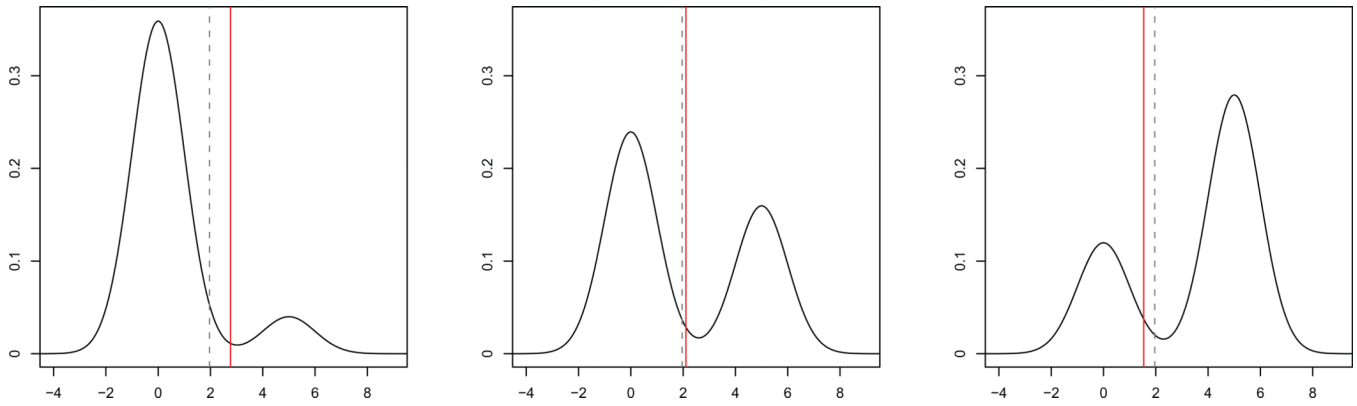

Figure 1.

Observed *q*-value thresholds for mixture distributions of the form $\pi_0 N(0, 1) + (1 - \pi_0) N(5, 1)$. For π_0=0.9 (left), 0.6 (center), and 0.3 (right), we simulated 50000 independent test statistics from the mixture distribution, referred the statistics to the standard normal distribution to obtain 50000 *p*-values, and transformed these to *q*-values by the method of Storey and Tibshirani (2003). This was repeated 101 times for each mixture, and the median thresholds for *q*-value 0.05 (variability across the 101 replications was very low) are shown by the solid vertical lines. The dashed lines represent the uncorrected threshold.

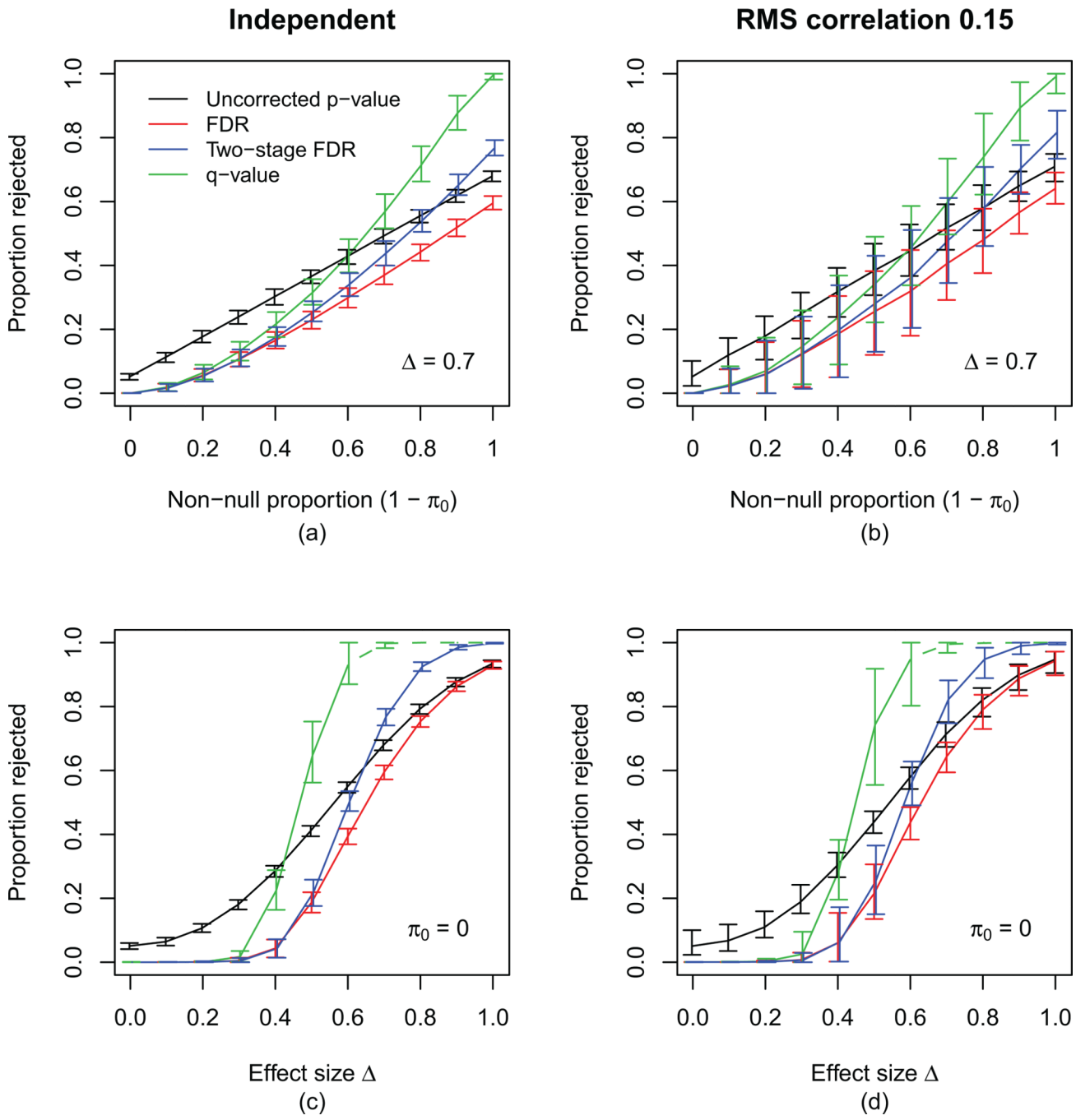

Figure 2.

Comparing two samples of size 25, where the first sample has distribution $N(0, 1)$ for each of 1000 variables, while the second sample has $1000\pi_0$ variables distributed as $N(0, 1)$ and the remaining $1000(1 - \pi_0)$ distributed as $N(\Delta, 1)$. The curves join the mean proportions of hypotheses rejected by each method, while the bars display the 5th and 95th percentiles, over 299 simulations. Upper subfigures: $\alpha = 0.7$, $\pi_0 = 0, 0.1, \ldots, 1$; results with (a) mutually independent variables, and (b) root mean square correlation 0.15. Lower subfigures: $\pi_0 = 0$, $\Delta = 0, 0.1, \ldots, 1$; (c) mutually independent variables, and (d) root mean square correlation 0.15. The dashed portions of the *q*-value curves indicate that for $\Delta > 0.6$, the algorithm produced an error for some simulations, due to a negative estimate of π_0. In the

independent-test simulations, this occurred for 1.3%, 17%, 54%, and 79% of the simulations with $\Delta = 0.7, 0.8, 0.9, 1.0$, respectively; for the simulations with RMS correlation 0.15, the corresponding values were 9.4%, 43%, 70%, and 85%.

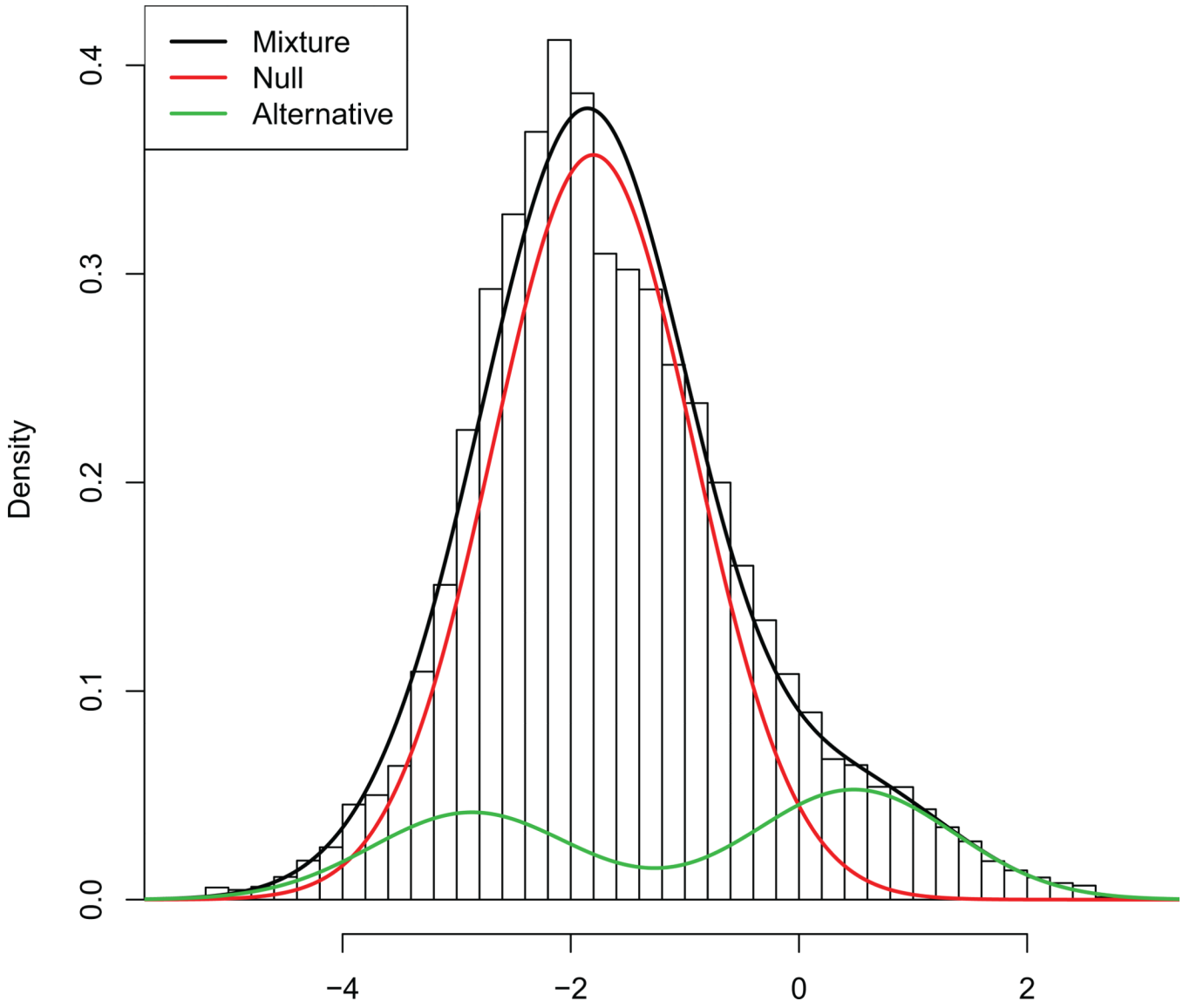

Figure 3.
Histogram of test statistics (Wald *t*-statistics transformed to *z*-scores) for cortical thickness at the 81924 vertices, from the study ofProal et al. (2011). Negative values imply diminished cortical thickness in individuals with ADHD. The density curves are generated by the FDR estimation method of Muralidharan (2009, 2010), which models the test statistic density as a mixture of an empirical null component (Efron, 2004, 2010) and an alternative component.

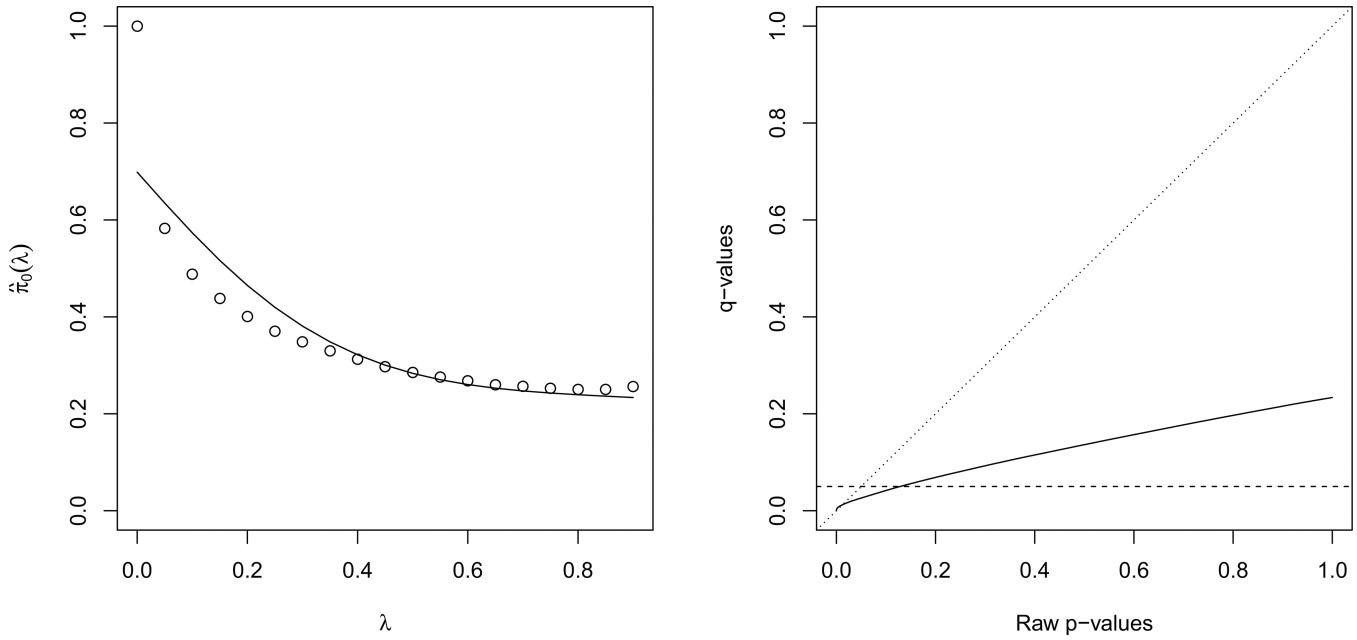

Figure 4.
Left: Cubic spline smooth of the null proportion estimates $\hat{\pi}_0(\lambda)$ (see (A.1)), for the cortical thickness data ofProal et al. (2011). Right: The ordered raw *p*-values plotted against the *q*-values given by (4); also shown are the line of identity and the horizontal line at *q*-value=0.05.

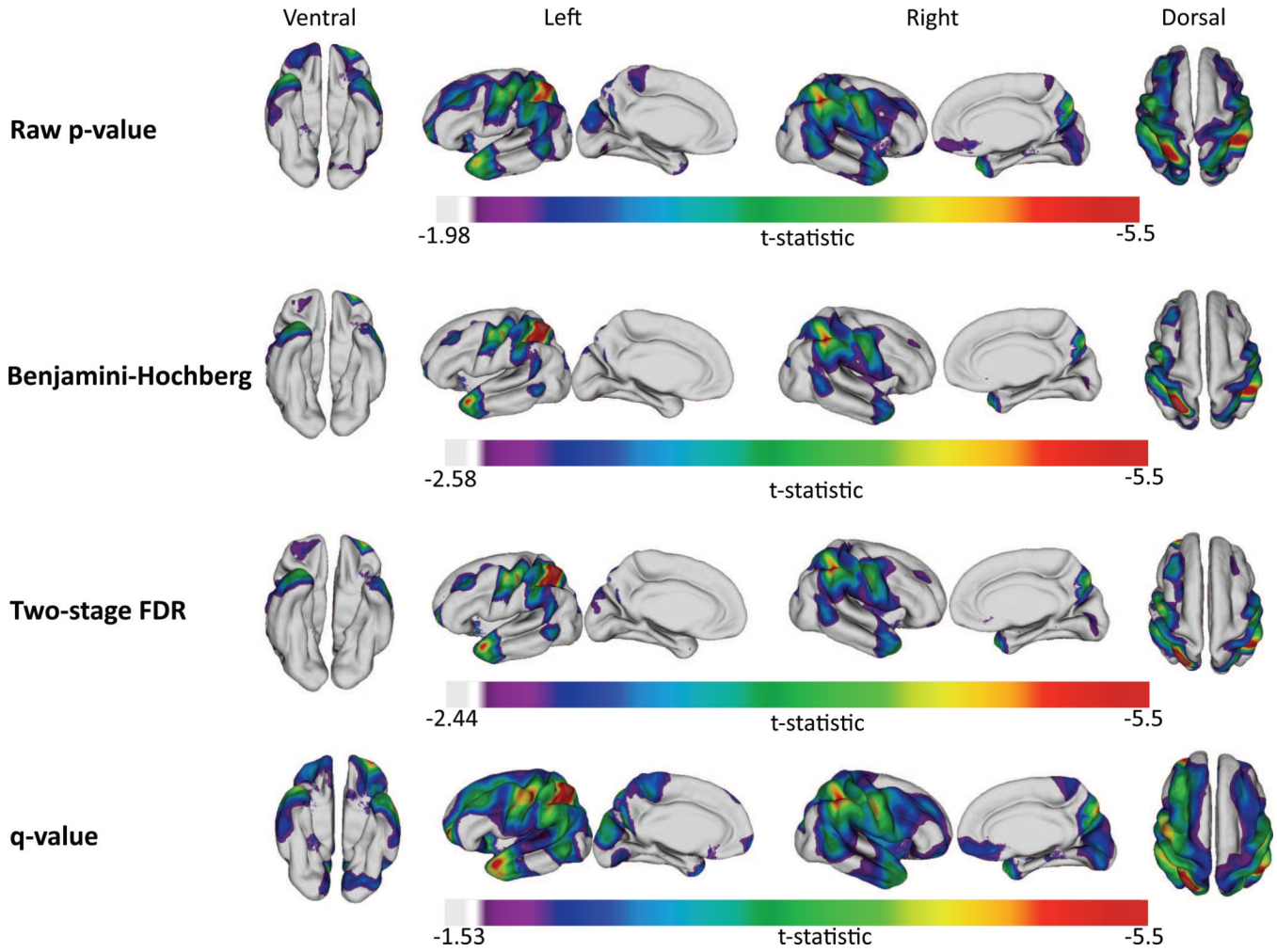

Figure 5.

Regions of significantly lower cortical thickness in the ADHD group, at level 0.05 of the error rates controlled, respectively, by the four multiple testing procedures: type-I error probability for raw *p*-values, and FDR for the other procedures. Note that the *t*-statistic magnitude threshold for rejection is lower for the *q*-value method (1.53) than for raw *p*-values (1.98), resulting in more extensive declarations of significance.

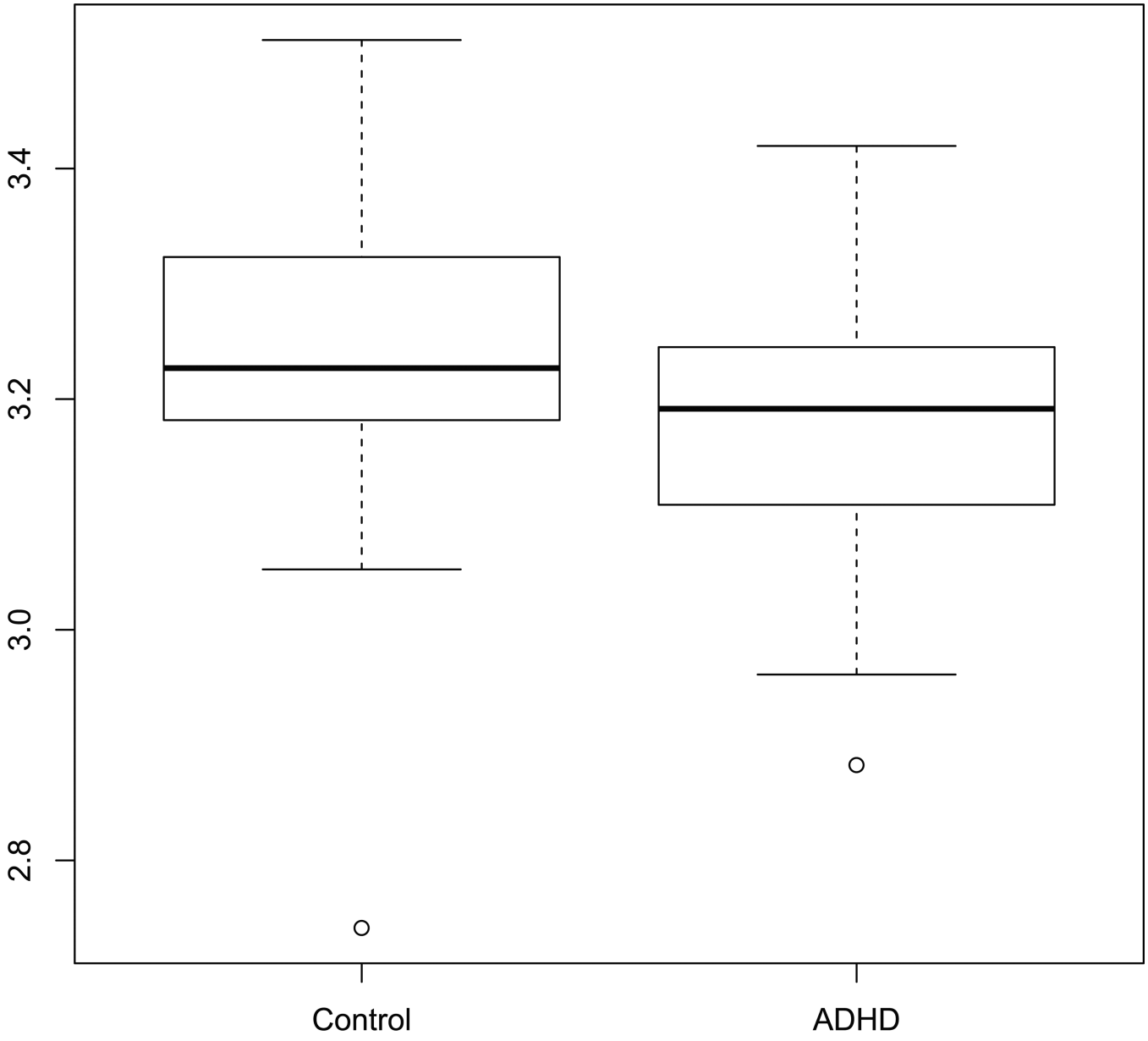

Figure 6.
Box plots of mean cortical thickness, over all 81924 vertices, for the 80 controls and the 59 individuals with ADHD.