*Statistical Methods in Medical Research* 2003; **12**: 419–446

# Controlling the familywise error rate in functional neuroimaging: a comparative review

**Thomas Nichols** and **Satoru Hayasaka** Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Functional neuroimaging data embodies a massive multiple testing problem, where 100 000 correlated test statistics must be assessed. The familywise error rate, the chance of any false positives is the standard measure of Type I errors in multiple testing. In this paper we review and evaluate three approaches to thresholding images of test statistics: Bonferroni, random field and the permutation test. Owing to recent developments, improved Bonferroni procedures, such as Hochberg's methods, are now applicable to dependent data. Continuous random field methods use the smoothness of the image to adapt to the severity of the multiple testing problem. Also, increased computing power has made both permutation and bootstrap methods applicable to functional neuroimaging. We evaluate these approaches on $t$ images using simulations and a collection of real datasets. We find that Bonferroni-related tests offer little improvement over Bonferroni, while the permutation method offers substantial improvement over the random field method for low smoothness and low degrees of freedom. We also show the limitations of trying to find an equivalent number of independent tests for an image of correlated test statistics.

## 1 Introduction

Functional neuroimaging refers to an array of technologies used to measure neuronal activity in the living brain. Two widely used methods, positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), both use blood flow as an indirect measure of brain activity. An experimenter images a subject repeatedly under different cognitive states and typically fits a massively univariate model. That is, a univariate model is independently fit at each of hundreds of thousands of volume elements, or *voxels*. Images of statistics are created that assess evidence for an experimental effect. Naive thresholding of 100 000 voxels at $\alpha = 5\%$ threshold is inappropriate, since 5000 false positives would be expected in null data.

False positives must be controlled over all tests, but there is no single measure of Type I error in multiple testing problems. The standard measure is the chance of any Type I errors, the familywise error rate (FWE). A relatively new development is the false discovery rate (FDR) error metric, the expected proportion of rejected hypotheses that are false positives. FDR-controlling procedures are more powerful then FWE procedures, yet still control false positives in a useful manner. We predict that FDR may soon eclipse FWE as the most common multiple false positive measure. In light of this, we believe that this is a choice moment to review FWE-controlling measures. (We prefer the

Address for correspondence: Thomas Nichols, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. E-mail: nichols@umich.edu

© Arnold 2003 10.1191/0962280203sm341ra

term *multiple testing problem* over *multiple comparisons problem*. 'Multiple comparisons' can allude to pairwise comparisons on a single model, whereas in imaging a large collection of models is each subjected to a hypothesis test.)

In this paper we attempt to describe and evaluate all FWE multiple testing procedures useful for functional neuroimaging. Owing to the spatial dependence of functional neuroimaging data, there are actually quite a small number of applicable methods. The only methods that are appropriate under these conditions are Bonferroni, random field methods and resampling methods. We limit our attention to finding FWE-corrected thresholds and *P*-values for Gaussian $Z$ and Student $t$ images. (An $\alpha_0$ FWE-corrected threshold is one that controls the FWE at $\alpha_0$, while an FWE-corrected *P*-value is the most significant $\alpha_0$ corrected threshold such that a test can be rejected.) In particular we do not consider inference on size of contiguous suprathreshold regions or clusters. We focus particular attention on low degrees-of-freedom (DF) $t$ images, as these are unfortunately common in group analyses. The typical images have tens to hundreds of thousands of tests, where the tests have complicated, though usually positive dependence structure.

In 1987 Hochberg and Tamhane wrote: 'The question of which error rate to control ... has generated much discussion in the literature'.[1] While that could be true in the statistics literature in the decades before their book was published, the same cannot be said of neuroimaging literature. When multiple testing has been acknowledged at all, the familywise error rate has usually been assumed implicitly. For example, of the two papers that introduced corrected inference methods to neuroimaging, only one explicitly mentions familywise error.[2,3] It is hoped that the introduction of FDR will enrich the discussion of multiple testing issues in the neuroimaging literature.

The remainder of the paper is organized as follows. We first introduce the general multiple comparison problem and FWE. We then review Bonferroni-related methods, followed by random field methods and then resampling methods. We then evaluate these different methods with 11 real datasets and simulations.

## 2 Multiple testing background in functional neuroimaging

In this section we formally define strong and weak control of familywise error, as well as other measures of false positives in multiple testing. We describe the relationship between the maximum statistic and FWE and also introduce step-up and step-down tests and the notion of equivalent number of independent tests.

### 2.1 Notation

Consider image data on a two- (2D) or three-dimensional (3D) lattice. Usually the lattice will be regular, but it may also be a irregular, corresponding to a 2D surface. Through some modeling process we have an image of test statistics $T = \{T_i\}$. Let $T_i$ be the value of the statistic image at spatial location $i$, $i \in \mathcal{V} = \{1, \ldots, V\}$, where $V$ is the number of voxels in the brain. Let the $H = H_i$ be a hypothesis image, where $H_i = 0$ indicates that the null hypothesis holds at voxel $i$, and $H_i = 1$ indicates that the alternative hypothesis holds. Let $H_0$ indicate the complete null case where $H_i = 0$ for all $i$. A decision to reject the null for voxel $i$ will be written $\hat{H}_i = 1$, not rejecting $\hat{H}_i = 0$.

Write the null distribution of $T_i$ as $F_{0,T_i}$, and let the image of $P$-values be $P = \{P_i\}$. We require nothing about the modeling process or statistic other than the test being unbiased, and, for simplicity of exposition, we will assume that all distributions are continuous.

## 2.2 Measures of false positives in multiple testing

A valid $\alpha$-level test at location $i$ corresponds to a rejection threshold $u$ where $\mathsf{P}\{T_i \geq u | H_i = 0\} \leq \alpha$. The challenge of the multiple testing problems to find a threshold $u$ that controls some measure of false positives across the entire image. While FWE is the focus of this paper, it is useful to compare its definition with other measures. To this end consider the cross-classification of all voxels in a threshold statistic image of Table 1.

In Table 1 $V$ is the total number of voxels tested, $V_{\cdot|1} = \sum_i H_i$ is the number of voxels with a false null hypothesis, $V_{\cdot|0} = V - \sum_i H_i$ the number true nulls, and $V_{1|\cdot} = \sum_i \hat{H}_i$ is the number of voxels above threshold, $V_{0|\cdot} = V - \sum_i \hat{H}_i$ the number below; $V_{1|1}$ is the number of suprathreshold voxels correctly classified as signal, $V_{1|0}$ the number incorrectly classified as signal; $V_{0|0}$ is the number of subthreshold voxels correctly classified as noise, and $V_{0|1}$ the number incorrectly classified as noise.

With this notation a range of false positive measures can be defined, as shown in Table 2. An observed familywise error (oFWE) occurs whenever $V_{1|0}$ is greater than zero, and the familywise error rate (FWE) is defined as the probability of this event. The observed FDR (oFDR) is the proportion of rejected voxels that have been falsely rejected, and the expected false discovery rate (FDR) is defined as the expected value of oFDR. Note the contradictory notation: an 'observed familywise error' and the 'observed false discovery rate' are actually *unobservable*, since they require knowledge of which tests are falsely rejected. The measures actually controlled, FWE and FDR, are the probability and expectation of the unobservable oFWE and oFDR respectively. Other variants on FDR have been proposed, including the positive FDR (pFDR), the FDR conditional on at least one rejection[4] and controlling the oFDR at some specified level of confidence (FDRc).[5] The per-comparison error rate (PCE) is essentially the 'uncorrected' measure of false positive, which makes no accommodation for multiplicity, while the per-family error rate (PFE) measures the expected count of false positives. Per-family errors can also be controlled with some level confidence (PFEc).[6] This taxonomy demonstrates that there are many potentially useful multiple false positive metrics, but we are choosing to focus on but one.

**Table 1** Cross-classification of voxels in a threshold statistic image

| | Null not rejected (declared inactive) | Null rejected (declared active) | |
|---|---|---|---|
| Null true (inactive) | $V_{0\|0}$ | $V_{1\|0}$ | $V - V_{\cdot\|1}$ |
| Null false (active) | $V_{0\|1}$ | $V_{1\|1}$ | $V_{\cdot\|1}$ |
| | $V - V_{1\|\cdot}$ | $V_{1\|\cdot}$ | $V$ |

**Table 2** Different measures of false positives in the multiple testing problem. $I_{\{A\}}$ is the indicator function for event $A$

| Measure of false positives | Abbreviation | Definition |
|---|---|---|
| Observed familywise error | oFWE | $V_{1\|0} > 0$ |
| Familywise error rate | FWE | P{oFWE} |
| Observed false discovery rate | oFDR | $V_{1\|0}/V_{1\|\cdot}\, I_{\{V_{1\|\cdot}>0\}}$ |
| Expected false discovery rate | FDR | E{oFDR} |
| Positive false discovery rate | pFDR | $E\{oFDR \mid V_{1\|\cdot} > 0\}$ |
| False discovery rate confidence | FDRc | $P\{oFDR \leq a\}$ |
| Per-comparison error rate | PCE | $E\{V_{1\|0}\}/V$ |
| Per-family error rate | PFE | $E\{V_{1\|0}\}$ |
| Per-family error rate confidence | PFEc | $P\{V_{1\|0} \leq a\}$ |

There are two senses of FWE control, weak and strong. Weak control of FWE only requires that false positives are controlled under the complete null $H_0$:

$$\mathrm{P}\left(\bigcup_{i\in\mathcal{V}}\{T_i \geq u\}\,\middle|\, H_0\right) \leq \alpha_0 \tag{1}$$

where $\alpha_0$ is the nominal FWE. Strong control requires that false positives are controlled for any subset $\mathcal{V}_0 \subset \mathcal{V}$ where the null hypothesis holds

$$\mathrm{P}\left(\bigcup_{i\in\mathcal{V}_0}\{T_i \geq u\}\,\middle|\, H_i = 0, i \in \mathcal{V}_0\right) \leq \alpha_0 \tag{2}$$

Significance determined by a method with weak control only implies that $H_0$ is false, and does not allow the localization of individual significant voxels. Because of this, tests with only weak control are sometimes called 'omnibus' tests. Significance obtained by a method with strong control allows rejection of individual $H_i$s while controlling the FWE at all nonsignificant voxels. This localization is essential to neuroimaging, and in the rest of this document we focus on strong control of FWE and we will omit the qualifier unless needed.

Note that variable thresholds, $u_i$, could be used instead of a common threshold $u$. Any collection of thresholds $\{u_i\}$ can be used as long as the overall FWE is controlled. Also note that FDR controls FWE weakly. Under the complete null, oFDR becomes an indicator for an oFWE, and the expected oFDR exactly the probability of an oFWE.

## 2.3 The maximum statistic and FWE

The maximum statistic, $M_T = \max_i T_i$, plays a key role in FWE control. The connection is that one or more voxels will exceed a threshold if and only if the maximum exceeds the threshold

$$\bigcup_i \{T_i \geq u\} = \{M_T \geq u\} \tag{3}$$

This relationship can be used to directly obtain a FWE-controlling threshold from the distribution of the maximum statistic under $H_0$. To control FWE at $\alpha_0$ let $u_{\alpha_0} = F^{-1}_{M_T|H_0}(1-\alpha_0)$, the $(1-\alpha_0)100$th percentile of the maximum distribution under the complete null hypothesis. Then $u_{\alpha_0}$ has weak control of FWE:

$$\begin{aligned}\mathrm{P}\left(\bigcup_i \{T_i \geq u_{\alpha_0}\} \middle| H_0\right) &= \mathrm{P}(M_T \geq u_{\alpha_0} | H_0) \\ &= 1 - F_{M_T|H_0}(u_{\alpha_0}) \\ &= \alpha_0\end{aligned}$$

Further, this $u_{\alpha_0}$ also has strong control of FWE, although an assumption of subset pivotality is required.[7] A family of tests has subset pivotality if the null distribution of a subset of voxels does not depend on the state of other null hypotheses. Strong control follows

$$\mathrm{P}\left(\bigcup_{i\in\mathcal{V}_0} \{T_i \geq u_{\alpha_0}\} \middle| H_i = 0,\, i \in \mathcal{V}_0\right) = \mathrm{P}\left(\bigcup_{i\in\mathcal{V}_0} \{T_i \geq u_{\alpha_0}\} \middle| H_0\right) \tag{4}$$

$$\leq \mathrm{P}\left(\bigcup_{i\in\mathcal{V}} \{T_i \geq u_{\alpha_0}\} \middle| H_0\right) \tag{5}$$

$$= \alpha_0 \tag{6}$$

where the first equality uses the assumption of subset pivotality. In imaging, subset pivotality is trivially satisfied, as the image of hypotheses $H$ satisfies the free combination condition. That is, there are no logical constraints between different voxel's null hypotheses, and all combinations of voxel level nulls ($\{0,1\}^V$) are possible. Situations where subset pivotality can fail include tests of elements of a correlation matrix (Ref. 7, p. 43).

Note that we could equivalently find $P$-value thresholds using the distribution of the minimum $P$-value $N_P = \min_i P_i$. Whether with $M_T$ or $N_P$, we stress that we are not simply making inference on the extremal statistic, but rather using its distribution to find a threshold that controls FWE strongly.

## 2.4 Step-up and step-down tests

A generalization of the single threshold test takes the form of multi-step tests. Multi-step tests consider the sequence of sorted $P$-values and compare each $P$-value to a different threshold. Let the ordered $P$-values be $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(V)}$ and $H_{(i)}$ be the null hypothesis corresponding to $P_{(i)}$. Each $P$-value is assessed according to

$$P_{(i)} \leq u_i \tag{7}$$

Usually $u_1$ will correspond to a standard fixed threshold test (e.g., Bonferroni $u_1 = \alpha_0/V$).

There are two types of multi-step tests, step-up and step-down. A step-up test proceeds from the least to most significant $P$-value ($P_{(V)}$, $P_{(V-1)}$, ...) and successively applies equation (7). The first $i'$ that satisfies (7) implies that all smaller $P$-values are significant; that is, $\hat{H}_{(i)} = 1$ for all $i \leq i'$, $\hat{H}_{(i)} = 0$ otherwise. A step-down test proceeds from the most to least significant $P$-value ($P_{(1)}$, $P_{(2)}$, ...). The first $i'$ that *does not* satisfy (7) implies that all strictly smaller $P$-values are significant; that is, $\hat{H}_{(i)} = 1$ for all $i < i'$, $\hat{H}_{(i)} = 0$ otherwise. For the same critical values $\{u_i\}$, a step-up test will be as or more powerful than a step-down test.

## 2.5 Equivalent number of independent tests

The main challenge in FWE control is dealing with dependence. Under independence, the maximum null distribution $F_{M_T|H_0}$ is easily derived:

$$F_{M_T}(t) = \prod_i F_{T_i}(t) = F_{T_1}^V(t) \tag{8}$$

where we have suppressed the $H_0$ subscript, and the last equality comes from assuming a common null distribution for all $V$ voxels. However, in neuroimaging data, independence is rarely a tenable assumption, as there is usually some form of spatial dependence either due to acquisition physics or physiology. In such cases the maximum distribution will not be known or even have a closed form. The methods described in this paper can be seen as different approaches to bounding or approximating the maximum distribution.

One approach that has an intuitive appeal, but which has not been a fruitful avenue of research, is to find an equivalent number of independent observations. That is, to find a multiplier $\theta$ such that

$$F_{M_T}(t) = F_{T_1}^{\theta V}(t) \tag{9}$$

or, in terms of $P$-values,

$$\begin{aligned} F_{N_P}(t) &= 1 - (1 - F_{P_{(1)}}(t))^{\theta V} \\ &= 1 - (1 - t)^{\theta V} \end{aligned} \tag{10}$$

where the second equality comes from the uniform distribution of $P$-values under the null hypothesis. We are drawn to the second form, for $P$-values, because of its simplicity and the log-linearity of $1 - F_{N_P}(t)$.

For the simulations considered below, we will assess if minimum $P$-value distributions follow equation (10) for a small $t$. If they do, one can find the effective number of tests $\theta V$. This could be called the number of maximum-equivalent elements in the data, or *maxels*, where a single maxel would consist of $V/(\theta V) = \theta^{-1}$ voxels.

## 3 FWE methods for functional neuroimaging

There are two broad classes of FWE methods, those based on the Bonferroni inequality and those based on the maximum statistic (or minimum *P*-value) distribution. We first describe Bonferroni-type methods and then two types of maximum distribution methods: random field theory-based and resampling-based methods.

### 3.1 Bonferroni and related

The most widely known multiple testing procedure is the 'Bonferroni correction'. It is based on the Bonferroni inequality, a truncation of Boole's formula.[1] We write the Bonferroni inequality as

$$\mathsf{P}\left\{\bigcup_i A_i\right\} \leq \sum_i \mathsf{P}\{A_i\} \tag{11}$$

where $A_i$ corresponds to the event of test $i$ rejecting the null hypothesis when true. The inequality makes no assumption on dependence between tests, although it can be quite conservative for strong dependence. As an extreme, consider that $V$ tests with perfect dependence ($T_i = T_{i'}$ for $i \neq i'$) require no correction at all for multiple testing.

However, for many independent tests the Bonferroni inequality is quite tight for typical $\alpha_0$. For example, for $V = 32^3$ independent voxels and $\alpha_0 = 0.05$, the exact one-sided *P*-value threshold is $1 - ((1 - \alpha_0)^{1/V}) = 1.5653 \times 10^{-6}$ while Bonferroni gives $\alpha_0/V = 1.5259 \times 10^{-6}$. For a $t_9$ distribution, this is the difference between 10.1616 and 10.1928. Surprisingly, for weakly-dependent data, Bonferroni can also be fairly tight. To preview the results below, for $32^3$-voxel $t_9$ statistic image based on Gaussian data with isotropic FWHM smoothness of three voxels, we find that the correct threshold is 10.0209. Hence Bonferroni thresholds and *P*-values can indeed be useful in imaging.

Considering another term of Boole's formula yields a second-order Bonferroni, or the Kounias inequality[1]

$$\mathsf{P}\left\{\bigcup_i A_i\right\} \leq \sum_i \mathsf{P}\{A_i\} - \max_{k=1,\ldots,V}\left\{\sum_{i \neq k} \mathsf{P}\{A_i \cap A_k\}\right\} \tag{12}$$

The Slepian or Dunn–Šidák inequalities can be used to replace the bivariate probabilities with products. The Slepian inequality, also known as the positive orthant dependence property, is used when $A_i$ corresponds to a one-sided test – it requires some form of positive dependence, like Gaussian data with positive correlations.[8] Dunn–Šidák is used for two-sided tests and is a weaker condition, for example, only requiring the data follow a multivariate Gaussian, t or F distribution (Ref. 7, p. 45).

If all the null distributions are identical and the appropriate inequality holds (Slepian or Dunn–Šidák), the second-order Bonferroni *P*-value threshold is $c$ such that

$$Vc - (V - 1)c^2 = \alpha_0 \tag{13}$$

When $V$ is large, however, $c$ will have to be quite small making $c^2$ negligible, essentially reducing to the first-order Bonferroni. For the example considered above, with $V = 32^3$ and $\alpha_0 = 0.05$, the Bonferroni and Kounias $P$-value thresholds agree to five decimal places ($0.05/V = 1.525881 \times 10^{-6}$ versus $c = 1.525879 \times 10^{-6}$).

Other approaches to extending the Bonferroni method are step-up or step-down tests. A Bonferroni step-down test can be motivated as follows. If we compare the smallest $P$-value $P_{(1)}$ to $\alpha_0/V$ and reject, then our multiple testing problem has been reduced by one test, and we should compare the next smallest $P$-value $P_{(2)}$ to $\alpha_0/(V-1)$. In general, we have

$$P_{(i)} \leq \alpha_0 \frac{1}{V - i + 1} \tag{14}$$

This yields the Holm step-down test,[9] which starts at $i = 1$ and stops as soon as the inequality is violated, rejecting all tests with smaller $P$-values. Using the very same critical values, the Hochberg step-up test[10] starts at $i = V$ and stops as soon as the inequality is satisfied, rejecting all tests with smaller $P$-values. However, the Hochberg test depends on a result of Simes.

Simes[11] proposed a step-up method that only controlled FWE weakly and was only proven to be valid under independence. In their seminal paper, Benjamini and Hochberg[12] showed that Simes' method controlled what they named the 'False Discovery Rate'. Both Simes' test and Benjamini and Hochberg's FDR have the form

$$P_{(i)} \leq \alpha_0 \frac{i}{V} \tag{15}$$

Both are step-up tests, which work from $i = V$.

The Holm method, like Bonferroni, makes no assumption on the dependence of the tests. But if the Slepian or Dunn–Šidák inequality holds the 'Šidák improvement on Holm' can be used.[13] The Šidák method is also a step-down test but uses thresholds $u_i = 1 - (1 - \alpha_0)^{1/(V-i+1)}$ instead.

Recently Benjamini and Yekutieli[14] showed that the Simes/FDR method is valid under 'positive regression dependency on subsets' (PRDS). As with Slepian inequality, Gaussian data with positive correlations will satisfy the PRDS condition, but it is more lenient than other results in that it only requires positive dependency on the null, that is, only between test statistics for which $H_i = 0$. Interestingly, since Hochberg's method depended on Simes' result, so Ref. 14 also implies that Hochberg's step-up method is valid under dependence.

Table 3 summarizes the multi-step methods. The Hochberg step-up method and Šidák step-down method appear to be the most powerful Bonferroni-related FWE procedures available for functional neuroimaging. Hochberg uses the same critical values as Holm, but Hochberg can only be more powerful since it is a step-up test. The Šidák has slightly more lenient critical values, but may be more conservative than Hochberg because it is a step-down method. If positive dependence cannot be assumed for one-sided tests, Holm's step-down method would be the best alternative. The Simes/FDR procedure has the most lenient critical values, but only controls FWE

**Table 3** Summary of multi-step procedures

| Procedure | $u_i$ | Type | FWE control | Assumptions |
|---|---|---|---|---|
| Holm | $\alpha_0(1/V - i + 1)$ | Step-down | Strong | None |
| Šidák | $1 - (1 - \alpha_0)^{1/(V - i + 1)}$ | Step-down | Strong | Slepian/Dunn–Šidák |
| Hochberg | $\alpha_0(1/V - i + 1)$ | Step-up | Strong | PRDS |
| Simes/FDR | $\alpha_0(i/V)$ | Step-up | Weak | PRDS |

weakly. See works by Sarkar[15,16] for a more detailed overview of recent developments in multiple testing and the interaction between FDR and FWE methods.

The multi-step methods can adapt to the signal present in the data, unlike Bonferroni. For the characteristics of neuroimaging data, with large images with sparse signals, however, we are not optimistic these methods will offer much improvement over Bonferroni. For example, with $\alpha_0 = 0.05$ and $V = 32^3$, consider a case where 3276 voxels (10%) of the image have a very strong signal, that is, infinitesimal *P*-values. For the 3276th smallest *P*-value the relevant critical value would be $0.05/(V - 3276 + 1) = 1.6953 \times 10^{-6}$ for Hochberg and $1 - (1 - 0.05)^{1/(V-3276+1)} = 1.7392 \times 10^{-6}$ for Šidák, each only slightly more generous than the Bonferroni threshold of $1.5259 \times 10^{-6}$ and a decrease of less than 0.16 in $t_9$ units. For more typical, even more sparse signals there will be less of a difference. (Note that the Simes/FDR critical value would be $0.05 \times 3276/V = 0.005$, although with no strong FWE control.)

The strength of Bonferroni and related methods are their lack of assumptions or only weak assumptions on dependence. However, none of the methods makes use of the spatial structure of the data or the particular form of dependence. The following two methods explicitly account for image smoothness and dependence.

## 3.2 Random field theory

Both the random field theory (RFT) methods and the resampling-based methods account for dependence in the data, as captured by the maximum distribution. The random field methods approximate the upper tail of the maximum distribution, the end needed to find small $\alpha_0$ FWE thresholds. In this section we review the general approach of the random field methods and specifically use the Gaussian random field results to give intuition to the methods. Instead of detailed formulae for every type of statistic image we motivate the general approach for thresholding a statistic image, and then briefly review important details of the theory and common misconceptions.

For a detailed description of the random field results we refer to a number of useful papers. The original paper introducing RFT to brain imaging[2] is a very readable and well-illustrated introduction to the Gaussian random field method. A later work[17] unifies Gaussian and *t*, $\chi^2$ and *F* field results. A very technical, but comprehensive summary[18] also includes results on Hotellings $T^2$ and correlation fields. As part of a broad review of statistical methods in neuroimaging, Ref. 19 describes RFT methods and highlights their limitations and assumptions.

### *3.2.1 RFT intuition*

Consider a continuous Gaussian random field, $Z(s)$ defined on $s \in \Omega \subset \mathbb{R}^D$, where $D$ is the dimension of the process, typically 2 or 3. Let $Z(s)$ have zero mean and unit variance, as would be required by the null hypothesis. For a threshold $u$ applied to the field, the regions above the threshold are known as the *excursion set*, $A_u = \Omega \cap \{s : Z(s) > u\}$. The *Euler characteristic* $\chi(A_u) \equiv \chi_u$ is a topological measure of the excursion set. While the precise definition involves the curvature of the boundary of the excursion set (Ref. 20, cited in Ref. 21), it essentially counts the number of connected suprathreshold regions or 'clusters', minus the number of 'holes' plus the number of 'hollows' (Figure 1). For high thresholds the excursion set will have no holes or hollows and $\chi_u$ will just count the number of clusters; for yet higher thresholds the $\chi_u$ will be either 0 or 1, an indicator of the presence of any clusters. This seemingly esoteric topological measure is actually very relevant for FWE control. If a null Gaussian statistic image $T$ approximates a continuous random field, then

$$\begin{aligned} \text{FWE} &= \mathsf{P}\{\text{oFWE}\} && (16)\\ &= \mathsf{P}\left\{\bigcup_i T_i \geq u\right\} && (17)\\ &= \mathsf{P}\{\max_i T_i \geq u\} && (18)\\ &\approx \mathsf{P}\{\chi_u > 0\} && (19)\\ &\approx \mathsf{E}\{\chi_u\} && (20) \end{aligned}$$

The first approximation comes from assuming that the threshold $u$ is high enough for there to be no holes or hollows and hence the $\chi_u$ is just counting clusters. The second approximation is obtained when $u$ is high enough such that the probability of two or more clusters is negligible.

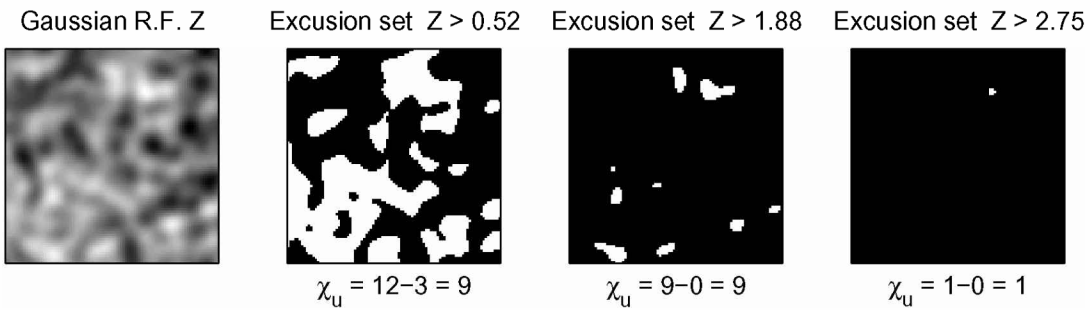


**Figure 1** Illustration of Euler characteristic ($\chi_u$) for different thresholds $u$. The left figure shows the random field, and the remaining images show the excursion set for different thresholds. The $\chi_u$ illustrated here only counts clusters minus holes (neglecting hollows). For high thresholds there are no holes, and $\chi_u$ just counts the number clusters.

The expected value of the $\chi_u$ has a closed-form approximation;[21] for $D = 3$

$$\mathsf{E}\{\chi_u\} \approx \lambda(\Omega)|\Lambda|^{1/2}(u^2 - 1)\exp(-u^2/2)/(2\pi)^2 \tag{21}$$

where $\lambda(\Omega)$ is the Lesbegue measure of the search region, the volume in three dimensions, and $\Lambda$ is the variance–covariance matrix of the gradient of the process,

$$\Lambda = \mathsf{Var}\left(\left[\frac{\partial}{\partial x}Z \quad \frac{\partial}{\partial y}Z \quad \frac{\partial}{\partial z}Z\right]^{\top}\right) \tag{22}$$

Its determinant $|\Lambda|$ is measure of roughness; the more 'wiggly' a process, the more variable the partial derivatives, the larger the determinant.

Consider an observed value $z$ of the process at some location. To build intuition consider the impact of search region size and smoothness on corrected $P$-value $P_z^c$. The corrected $P$-value is the upper tail area of the maximum distribution:

$$P_z^c = \mathsf{P}(\max_s Z(s) \geq z) \approx \mathsf{E}\{\chi_z\} \tag{23}$$

For large $z$, equation (21) gives

$$P_z^c \propto \lambda(\Omega)|\Lambda|^{1/2}z^2\exp\left(-\frac{z^2}{2}\right) \tag{24}$$

approximately. First note that, all other things constant, increasing large $z$ reduces the corrected $P$-value. Of course $P$-values must be nonincreasing in $z$, but note that equation (24) is not monotonic for all $z$, and that $\mathsf{E}\{\chi_z\}$ can be greater than 1 or even negative! Next observe that increasing the search region $\lambda(\Omega)$ increases the corrected $P$-value, decreasing significance. This should be anticipated, since an increased search volume results in a more severe multiple testing problem. And next consider the impact of smoothness, the inverse of roughness. Increasing smoothness decreases $|\Lambda|$, which in turn decreases the corrected $P$-value and increases significance. The intuition here is that an increase in smoothness reduces the severity of the multiple testing problem; in some sense there is less information with greater smoothness. In particular, consider that in the limit of infinite smoothness the entire processes has a common value, and there is no multiple testing problem at all.

### *3.2.2 RFT strengths and weaknesses*

As presented above, the merit of the RFT results are that they adapt to the volume searched (like Bonferroni) and to the smoothness of the image (unlike Bonferroni). When combined with the general linear model (GLM), the random field methods comprise a flexible framework for neuroimaging inference. For functional neuroimaging data that is intrinsically smooth (PET, SPECT, MEG or EEG) or heavily smoothed (multisubject fMRI), these results provide a unified framework to find FWE-corrected inferences for statistic images from a GLM. While we only discuss results for peak

statistic height, a family of available results includes *P*-values for the size of a cluster, the number of clusters and joint inference on peak height and cluster size.

Further, the results only depend on volume searched and the smoothness (see below for more details on edge corrections), and are not computationally burdensome. Finally, they have been incorporated into various neuroimaging software packages and are widely used (if poorly understood) by thousands of users. (The software packages include SPM, http://www.fil.ion.ucl.ac.uk; VoxBo, www.voxbo.org; FSL, www.fmrib.ox.ac.uk/fsl; and Worsley's own fmristat, http://www.math.mcgill.ca/keith/fmristat.)

The principal weakness of the random field methods are the assumptions. The assumptions are sometimes misstated, so we carefully itemize them.

1) The multivariate distribution of the image is Gaussian or derived from Gaussian data (e.g., for t or F statistic image).
2) The discretely sampled statistic images are assumed to be sufficiently smooth to approximate the behaviour of a continuous random field. The recommended rule of thumb is three voxels FWHM smoothness,[19] although we will critically assess this with simulations and data (see below for definition of FWHM smoothness).
3) The spatial autocorrelation function (ACF) must have two derivatives at the origin. Except for the joint cluster-peak height test,[22] the ACF is *not* assumed to have the form of a Gaussian density.
4) The data must be stationary or there must exist a deformation of space such that the transformed data is stationary.[23] This assumption is most questionable for reconstructed MEG and EEG data, which may have very convoluted covariance structures. Remarkably, for the peak height results we discuss here, nonstationarity need not even be estimated (see below for more.)
5) The results assume that roughness/smoothness is known with negligible error. Poline *et al.*[24] found that uncertainty in smoothness was in fact appreciable, causing corrected *P*-values to be accurate to only $\pm 20\%$ if smoothness was estimated from a single image. (In a recent abstract, Worsley proposed methods for nonstationary cluster size inference, which accounts for variability in smoothness estimation.[25])

### 3.2.3 *RFT essential details*

To simplify the presentation, the results above avoided several important details which we now review.

***Roughness and RESELS.*** Because the roughness parameter $|\Lambda|^{1/2}$ lacks interpretability, Worsley proposed a reparameterization in terms of the convolution of a white noise field into a random field with smoothness that matches the data.

Consider a white noise Gaussian random field convolved with a Gaussian kernel. If the kernel has variance matrix $\Sigma$ then the roughness of the convolved random field is $\Lambda = \Sigma^{-1}/2$.[26] If $\Sigma$ has $\sigma_x^2$, $\sigma_y^2$, $\sigma_z^2$ on the diagonal and zero elsewhere, then

$$|\Lambda|^{1/2} = (|\Sigma|^{-1} 2^{-3})^{1/2} \tag{25}$$

$$= (\sigma_x \sigma_y \sigma_z)^{-1} 2^{-3/2} \tag{26}$$

To parameterize in terms of smoothness,Worsley used the *full width at half maximum* (FWHM), a general measure of spread of a density. For a symmetric density $f$ centered about zero FWHM is the value $x$ such that $f(-x/2) = f(x/2) = f(0)/2$. A Gaussian kernel has a FWHM of $\sigma\sqrt{8\log 2}$. If a white noise field is convolved with a Gaussian kernel with scale ($\text{FWHM}_x$, $\text{FWHM}_y$, $\text{FWHM}_z$) (and zero correlations), the roughness parameter is

$$|\Lambda|^{1/2} = \frac{(4\log 2)^{3/2}}{\text{FWHM}_x \text{FWHM}_y \text{FWHM}_z} \tag{27}$$

Worsley defined a *RES*olution *EL*ement, or RESEL to be a spatial element with dimensions $\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z$. The denominator of (27) is then the volume of one RESEL.

Noting that $\mathsf{E}\{\chi_u\}$ in equation (21) depends on the volume and roughness through $\lambda(\Omega)|\Lambda|^{1/2}$, it can be seen that search volume and RESEL size can be combined and instead written as the search volume measured in RESELs:

$$R_3 = \frac{\lambda(\Omega)}{\text{FWHM}_x \text{FWHM}_y \text{FWHM}_z} \tag{28}$$

The RFT results then depend only on this single quantity, a resolution-adjusted search volume, the RESEL volume. The essential observation was that $|\Lambda|^{1/2}$ can be interpreted as a function of FWHM, the scale of a Gaussian kernel required to convolve a white noise field into one with the same smoothness as the data. When $|\Lambda|^{1/2}$ is unknown it is estimated from the data (see below), but *not* by assuming that the ACF is Gaussian. A Gaussian ACF is *not* assumed by random field results, rather Gaussian ACF is only used to reparameterize roughness into interpretable units of smoothness. If the true ACF is not Gaussian the accuracy of the resulting threshold is not impacted, only the precise interpretation of RESELs is disturbed.

***Component fields and smoothness estimation.*** For the Gaussian case presented above, the smoothness of the statistic image under the null hypothesis is the key parameter of interest. For results on other types of fields including $t$, $\chi^2$ and $F$, the smoothness parameter describes the smoothness of the *component fields*. If each voxel's data are fit with a general linear model $Y = X\beta + \epsilon$, the component fields are images of $\epsilon_j/\sqrt{\mathsf{Var}\{\epsilon_j\}}$, where $\epsilon_j$ is scan $j$'s error. That is, the component fields are the unobservable, mean zero, unit variance Gaussian noise images that underlie the observed data.

Estimation of component field smoothness is performed on standardized residual images,[27] not the statistic image itself. The statistic image is not used because it will generally contain signal, increasing roughness and decreasing estimated smoothness. Additionally, except for the $Z$ statistic, the smoothness of the null statistic image will be different from that of the component fields. For example, see Ref. 21, Equation (7) and Ref. 26, appendix G for the relationship between $t$ and component field smoothness.

***Edge corrections and unified RFT results.*** The original Gaussian RFT results (21) assumed a negligible chance of the excursion set $A_u$ touching the boundary of the search region $\Omega$.[21] If clusters did contact the surface of $\Omega$ they would have a contribution less

than unity to $\chi_u$. Worsley developed correction terms to (21) to account for this effect.[17,18] These 'unified' results have the form

$$P_u^c = \sum_{d=0}^{D} R_d \rho_d(u) \tag{29}$$

where $D$ is the number of dimensions of the data, $R_d$ is the $d$-dimensional RESEL measure and $\rho_d(u)$ is the Euler characteristic density. These results are convenient as they dissociate the terms that depend only on the topology of the search regions ($R_d$) from those that depend only on the type of statistical field ($\rho_d(u)$).

***Nonstationarity and cluster size tests.*** For inferences on peak height, with the appropriate estimator of average smoothness,[23] equation (21) will be accurate in the presence of nonstationarity or variable smoothness. However, cluster size inference is greatly effected by nonstationarity. In a null statistic image, large clusters will be more likely in smooth regions and small clusters more likely in rough regions. Hence an incorrect assumption of stationarity will likely lead to inflated false positive rate in smooth regions and reduced power in rough regions.

As alluded to above, the solution is to deform space until stationarity holds (if possible[29]). Explicit application of this transformation is actually not required, and local roughness can be used to determine cluster sizes in the transformed space.[23,25,30]

***RESEL Bonferroni.*** A common misconception is that the random field results apply a Bonferroni correction based on the RESEL volume.[31] They are actually quite different results. Using Mill's ratio, the Bonferroni corrected $P$-value can be seen to be approximately

$$P_{\text{Bonf}}^c \propto V e^{-u^2/2} u^{-1} \tag{30}$$

While the RFT $P$-value for 3D data is approximately

$$P_{\text{RFT}}^c \propto R_3 e^{-u^2/2} u^2 \tag{31}$$

where $R_3$ is the RESEL volume (28). Replacing $V$ with $R_3$ obviously does not align these two results, nor are they even proportional. We will characterize the performance of a naive RESEL Bonferroni approach in Section 4.

***Gaussianized t images.*** Early implementation of random field methods (e.g., SPM96 and previous versions) used Gaussian RFT results on $t$ images. While an image of $t$ statistics can be converted into $Z$ statistics with the probability integral transform, the resulting processes is not a $t$ random field. Worsley[17] found that the degrees of freedom would have to be quite high, as many as 120 for a $t$ field to behave like a Gaussian field. We will examine the performance of the Gaussianized $t$ approach with simulations.

### *3.2.4 RFT conclusion*

In this subsection we have tried to motivate the RFT results, as well as highlight important details of their application. They are a powerful set of tools for data that are

smooth enough to approximate continuous random fields. When the data are insufficiently smooth, or when other assumptions are dubious, nonparametric resampling techniques may be preferred.

## 3.3 Resampling methods for FWE control

The central purpose of the random field methods is to approximate the upper tail of the maximal distribution $F_{M_T}(t)$. Instead of making assumptions on the smoothness and distribution of the data, another approach is to use the data itself to obtain an empirical estimate of the maximal distribution. There are two general approaches, permutation-based and bootstrap-based. Excellent treatments of permutation tests[32,33] and the bootstrap[34,35] are available, so here we only briefly review the methods and focus on the differences between the two approaches and specific issues relevant to neuroimaging.

To summarize briefly, both permutation and bootstrap methods proceed by resampling the data under the null hypothesis. In the permutation test the data is resampled *without* replacement, while in a bootstrap test the residuals are resampled *with replacement* and a null dataset constituted. To preserve the spatial dependence the resampling is not performed independently voxel by voxel, but rather entire images are resampled as a whole. For each resampled null dataset, the model is fit, the statistic image computed, and the maximal statistic recorded. By repeating this process many times an empirical null distribution of maximum $\hat{F}_{M_T}$ is constructed, and the $100(1-\alpha_0)$th percentile provides an FWE-controlling threshold. For a more detailed introduction to the permutation test for functional neuroimaging see Ref. 36.

We now consider three issues that have an impact on the application of resampling methods to neuroimaging.

### *3.3.1 Voxel-level statistic and homogeneity of specificity and sensitivity*

The first issue we consider is common to all resampling methods used to find the maximum distribution and an FWE threshold. While a parametric method will assume a common null distribution for each voxel in the statistic image, $F_{0,T_i} = F_0$, FWE resampling methods are actually valid *regardless* of whether the null distributions $\{F_{0,T_i}\}$ are the same. This is because the maximum distribution captures any heterogeneity; as equation (4) shows, the relationship between the maximum and FWE makes no assumption of homogeneous voxel null distributions. Nonparametric methods will accurately reflect the maximum distribution of whatever statistic is chosen, and produce valid estimates of FWE-controlling thresholds.

However, once an FWE-controlling threshold is chosen, the false positive rate and power at each voxel depends on each voxel's distribution. As shown in Figure 2, FWE can be controlled overall, but if the voxel null distributions are heterogeneous, the Type I error rate will be higher at some voxels and lower at others. As a result, even though nonparametric methods admit the use of virtually any statistic (e.g., raw percent change, or mean difference), we prefer a voxel-level statistic that has a common null distribution $F_0$ (e.g., $t$ or $F$). Hence the usual statistics motivated by parametric theory are used to provide a more pivotal $T_i$ than an un-normalized statistic. Note that even though the statistic image may use a parametric statistic, the FWE-corrected $P$-values are nonparametric.

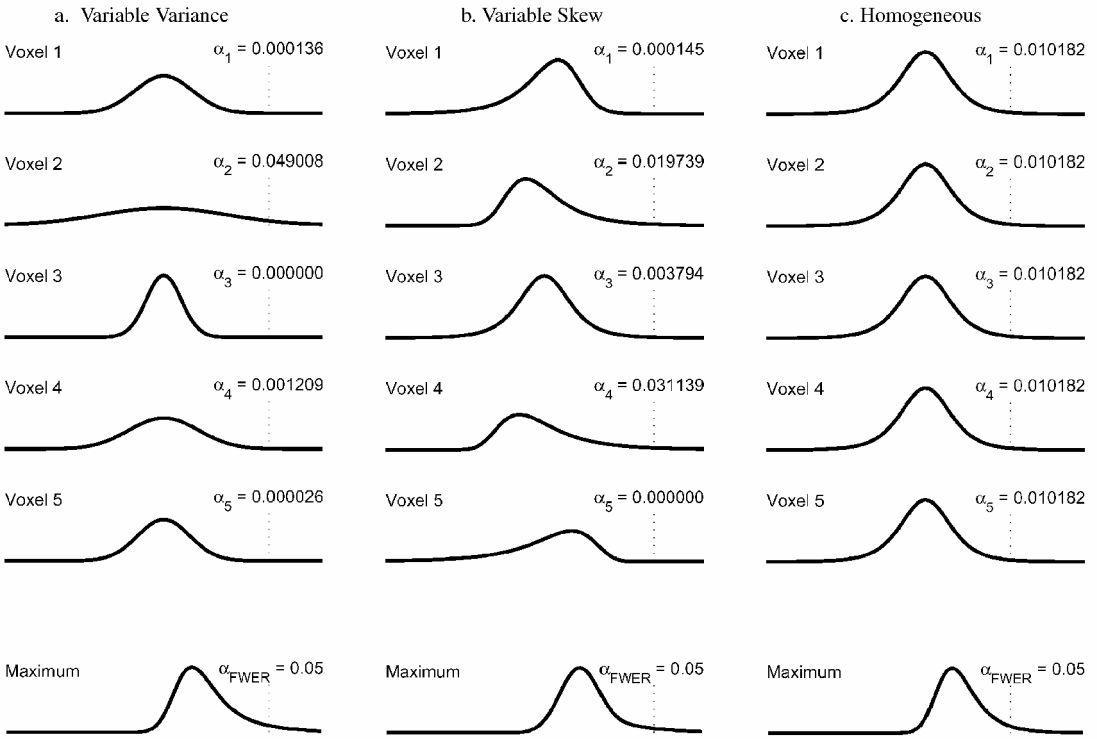


**Figure 2** Impact of heterogeneous null distributions on FWE control. Shown are the null distributions for five independent voxels, the null distribution of the maximum of the five voxels, and the 5% FWE thresholds. (a) Use of the mean difference statistic allows variance to vary from voxel to voxel, even under the null hypothesis. Voxel 2 has relatively large variance and shifts the maximum distribution to the right; the risk of Type I error is largely due to voxel 2, and, in contrast, voxel 3 will almost never generate a false positive. (b) If a *t* statistic is used the variance is standardized but the data may still exhibit variable skew. This would occur if the data are not Gaussian *and* have heterogeneous skew. Here voxels 2 and 4 bear most of the FWE risk. (c) If the voxel-level null distributions are homogeneous (e.g. if the *t* statistic is used and the data are Gaussian) there will be uniform risk of false positives. In all three of these cases the FWE is controlled, but the risk of Type I error may not be evenly distributed.

### *3.3.2 Randomization versus permutation versus bootstrap*

The permutation and bootstrap tests are most readily distinguished by their sampling schemes (without versus with replacement). However, there are several other important differences, and subtle aspects to the justification of the permutation test. These are summarized in Table 4.

A permutation test requires an assumption of exchangeability under the null hypothesis. This is typically justified by an assumption of independence and identical distribution. However, if a randomized design is used, no assumptions on the data are required at all. A *randomization test* uses the random selection of the experimental design to justify the resampling of the data (or, more precisely, the relabeling of the data). While the permutation test and randomization test are often referred to by the same name, we find it useful to distinguish them. As the randomization test supposes no

**Table 4** Differences between randomization, permutation and bootstrap tests

| | Randomization | Permutation | Bootstrap |
|---|---|---|---|
| *P*-values | Exact | Exact | Asymptotic/approximate |
| Assumption | Randomized experiment | Ho-exchangeability | I.I.D. |
| Inference | Sample only | Population | Population |
| Models | Simple | Simple | General |

population, the resulting inferences are specific to the sample at hand. The permutation test, in contrast, uses a population distribution to justify resampling and hence makes inference on the population sampled.[37]

A strength of randomization and permutation tests is that they exactly control the false positive rate. Bootstrap tests are only asymptotically exact, and each particular type of model should be assessed for specificity of the FWE thresholds. We are unaware of any studies of the accuracy of the bootstrap for the maximum distribution in functional neuroimaging. Further, the permutation test allows the slightly more general condition of exchangeability, in contrast to the bootstrap's independent and identically distributed assumption.

The clear advantage of the bootstrap is that it is a general modeling method. With the permutation test, each type of model must be studied to find the nature of exchangeability under the null hypothesis. And some data, such as positive one-sample data (i.e., not difference data) cannot be analysed with a permutation test, as the usual statistics are invariant to permutations of the data. The bootstrap can be implemented generally for a broad array of models. While we do not assert that bootstrap tests are automatic, and indeed general linear model design matrices can be found where the bootstrap performs horribly (see Ref. 35, p. 276), it is a more flexible approach than the permutation test.

### *3.3.3 Exchangeability and fMRI time series*

Functional magnetic resonance imaging (fMRI) data is currently the most widely used functional neuroimaging modality. However, fMRI time series exhibit temporal autocorrelation that violates the exchangeability/independence assumption of the resampling methods. Three strategies to deal with temporal dependence have been applied: do nothing, resample ignoring autocorrelation;[38] use a randomized design and randomization test;[39] and decorrelate and then resample.[40–42] Ignoring the autocorrelation in parametric settings tends to inflate significance due to biased variance estimates; with nonparametric analyses there may be either inflated or deflated significance depending on the resampling schemes. In a randomization test the data is considered fixed, and hence any autocorrelation is irrelevant to the validity of the test (power surely does depend on the autocorrelation and the resampling scheme, but this has not been studied to our knowledge). The preferred approach is the last one. The process consists of fitting a model, estimating the autocorrelation with the residuals, decorrelating the residuals, resampling, and then recorrelating the resampled residuals, creating null hypothesis realizations of the data. The challenges of this approach are the estimation of the autocorrelation and the computational burden of the decorrelation–recorrelation

process. To have an exact permutation test the residuals must be exactly whitened, but this is impossible without the true autocorrelation. However, in simulations and with real null data, Brammer and colleagues found that the false positive rate was well controlled.[40] To reduce the computational burden, Fadili and Bullmore[43] proposed performing the entire analysis in the whitened (i.e., wavelet) domain.

#### *3.3.4 Resampling conclusion*

Nonparametric permutation and bootstrap methods provide estimation of the maximum distribution without strong assumptions, and without inequalities that loosen with increasing dependence. Only their computational intensity and lack of generality preclude their widespread use.

## 4 Evaluation of FWE methods

We evaluated methods from the three classes of FWE-controlling procedures. Of particular interest is a comparison of random field and resampling methods, permutation in particular. In earlier work[36] comparing permutation and RFT methods on small group PET and fMRI data, we found the permutation method to be much more sensitive, and the RFT method comparable to Bonferroni. The present goal is to examine more datasets to see if those results generalize, and to examine simulations to discover if the RFT method is intrinsically conservative or if specific assumptions did not hold in the datasets considered. In particular, we seek the minimum smoothness required by the random field theory methods to perform well. We also investigate if two of the extended Bonferroni methods enhance the sensitivity of Bonferroni.

### 4.1 Real data results

We applied Bonferroni-related, random field and permutation methods to nine fMRI group datasets and two PET datasets. All data were analysed with a mixed effect model based on summary statistics.[44] This approach consists of fitting intrasubject general linear models on each subject, creating a contrast image of the effect of interest and assessing the population effect with a one-sample $t$ test on the contrast images. The smoothness parameter of the random field results were estimated from the standardized residual images of the one-sample $t$.[27] Random field results were obtained with SPM99 (http://www.fil.ion.ucl.ac.uk/spm) and nonparametric results were obtained with SnPM99 (http://www.fil.ion.ucl.ac.uk/spm/snpm).

A detailed description of each dataset is omitted for reasons of space, but we summarize each briefly. *Verbal Fluency* is a five-subject PET dataset comparing a baseline of passive listening versus word generation as cued by single letter (complete dataset available at http://www.fil.ion.ucl.ac.uk/spm/data). *Location Switching* and *Task Switching* are two different effects from a 10-subject fMRI study of attention switching (Tor Wager *et al.*, in preparation). *Faces: Main Effect* and *Faces: Interaction* are two effects (main effect data available at http://www.fil.ion.ucl.ac.uk/spm/data) from a 12-subject fMRI study of repetition priming.[45] *Item Recognition* is one effect from a 12-subject fMRI study of working memory.[46] *Visual Motion* is a 12-subject PET study of visual motion perception, comparing moving squares to fixed ones.[47]

*Emotional Pictures* is one effect from a 13-subject fMRI study of emotional processing, as probed by photographs of varying emotional intensity.[48] *Pain: Warming*, *Pain: Anticipation* and *Pain: Pain* are three effects from a 23-subject fMRI study of pain and the placebo effect (Tor Wager *et al.*, in preparation).

Tables 5 and 6 shows the results for the 11 datasets. Table 5 shows that for every dataset the permutation method has the lowest threshold, often dramatically so. Using either Bonferroni or permutation as a reference, the RFT becomes more conservative with decreasing degrees of freedom (DF), for example specifying a threshold of 4701.32 for a 4 DF analysis. The Bonferroni threshold is lower than the RFT threshold for all the low-DF studies. Only for the 22 DF study is the RFT threshold below Bonferroni, although the two approaches have comparable thresholds for one of the 11 DF studies and the 2 DF study. The smoothness is greater than three voxel FWHM for all studies, except for the *z*-smoothness in the visual motion study. This suggests that a three voxel FWHM rule of thumb[19] is insufficient for low-DF *t* statistic images.

Degrees of freedom and not smoothness seems to be the biggest factor in the convergence of the RFT and permutation results. That is, RFT comes closest to permutation *not* when smoothness is particularly large (e.g., *Task switching*), but when degrees of freedom exceed 12 (e.g., the *Pain*: dataset). This suggest a conservativeness in the low-DF RFT *t* results that is not explained by excessive roughness.

Comparing the 11 DF studies *Item recognition* and *Visual motion* is informative, as one has twice as many voxels and yet half as many RESELs. This situation results in Bonferroni being higher on *Item recognition* (9.80 versus 8.92) yet RFT being lower (9.87 versus 11.07). *Item recognition* has the lower permutation threshold (7.67 versus 8.40) suggesting that the resampling approach is adapting to greater smoothness despite the larger number of voxels.

Hochberg and Šidák are often infinity, indicating that no $\alpha_0 = 0.05$ threshold exists [i.e., no *P*-value satisfied equation (7)]. Also note that Hochberg and Šidák can be more

**Table 5** Summary of FWE inferences for 11 PET and fMRI studies. 5% FWE thresholds for five different methods are presented, RFT, Bonferroni, Hochberg step-up, Šidák step-down and permutation. Note how RFT only outperforms other methods for studies with the largest degrees of freedom. Hochberg and Šidák's method rarely differs from Bonferroni by much. Permutation always has the lowest threshold

| Study | DF | Voxels | Voxel FWHM smoothness | RESEL volume | FEW-corrected *t* threshold | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | RFT | Bonf. | Hoch. | Šidák | Perm. |
| Verbal fluency | 4 | 55027 | 5.6 6.3 3.9 | 399.9 | 4701.32 | 42.59 | ∞ | ∞ | 10.14 |
| Location switching | 9 | 36124 | 6.1 5.9 5.1 | 196.8 | 11.17 | 10.31 | ∞ | ∞ | 5.83 |
| Task switching | 9 | 37181 | 6.4 6.9 5.2 | 161.9 | 10.79 | 10.35 | 10.29 | 10.29 | 5.10 |
| Faces: main effect | 11 | 51560 | 4.1 4.1 4.3 | 713.3 | 10.43 | 9.07 | 9.07 | 9.04 | 7.92 |
| Faces: interaction | 11 | 51560 | 3.8 3.9 4.0 | 869.8 | 10.70 | 9.07 | ∞ | ∞ | 8.26 |
| Item recognition | 11 | 110776 | 5.1 6.8 6.9 | 462.9 | 9.87 | 9.80 | 9.99 | 9.99 | 7.67 |
| Visual motion | 11 | 43724 | 3.9 4.4 2.2 | 1158.2 | 11.07 | 8.92 | 8.90 | 8.87 | 8.40 |
| Emotional pictures | 12 | 44552 | 5.6 5.4 5.0 | 294.7 | 8.48 | 8.41 | ∞ | ∞ | 7.15 |
| Pain: warning | 22 | 23263 | 4.7 4.9 3.5 | 288.6 | 5.93 | 6.05 | 6.09 | 6.04 | 4.99 |
| Pain: anticipation | 22 | 23263 | 5.0 5.1 3.6 | 253.4 | 5.87 | 6.05 | 6.07 | 6.07 | 5.05 |
| Pain: pain | 22 | 23263 | 4.6 4.8 3.4 | 309.9 | 5.95 | 6.05 | 6.05 | 6.05 | 5.15 |

**Table 6** Summary of FWE inferences for 11 PET and fMRI studies (continued). Shown are the number of significant voxels detected with the five methods discussed, along with permutation method on the smoothed variance *t* statistic

| Study | Number of significant voxels | | | | | |
|---|---|---|---|---|---|---|
| | *t* | | | | | Sm. Var *t* Perm. |
| | RFT | Bonf. | Hoch. | Šidák | Perm. | |
| Verbal fluency | 0 | 0 | 0 | 0 | 0 | 0 |
| Location switching | 0 | 0 | 0 | 0 | 158 | 354 |
| Task switching | 4 | 6 | 7 | 7 | 2241 | 3447 |
| Faces: main effect | 127 | 371 | 372 | 379 | 917 | 4088 |
| Faces: interaction | 0 | 0 | 0 | 0 | 0 | 0 |
| Item recognition | 5 | 5 | 4 | 4 | 58 | 378 |
| Visual motion | 626 | 1260 | 1269 | 1281 | 1480 | 4064 |
| Emotional pictures | 0 | 0 | 0 | 0 | 0 | 7 |
| Pain: warning | 127 | 116 | 116 | 118 | 221 | 347 |
| Pain: anticipation | 74 | 55 | 55 | 55 | 182 | 402 |
| Pain: pain | 387 | 349 | 350 | 353 | 732 | 1300 |

stringent than Bonferroni even though their critical values $u_i$ are never less than $\alpha_0/V$. This occurs because the critical $P$-value falls below both $u_i$ and $\alpha_0/V$.

Table 6 shows how, even though the permutation thresholds are always lower, it fails to detect any voxels in some studies. (As noted in Ref. 45, the *Faces: Interaction* effect is significant in an *a priori* region of interest.) While truth is unknown here, this is evidence of permutation's specificity. The last column of this table includes results using a smoothed variance *t* statistic, a means to boost degrees of freedom by 'borrowing strength' from neighboring voxels.[49,36] In all of these studies it increased the number of detected voxels, in some cases dramatically.

## 4.2 Simulation methods

We simulated $32 \times 32 \times 32$ images, since a voxel count of $32^3 = 32\,767$ is typical for moderate resolution ($\approx 3$ mm$^3$) data. Smooth images were generated as Gaussian white noise convolved with a 3D isotropic Gaussian kernel of size 0, 1.5, 3, 6 and 12 voxels FWHM ($\sigma = 0$, 0.64, 1.27, 2.55, 5.1). We did not simulate in the Fourier domain to avoid wrap-around effects, and to avoid edge effects of the kernel we padded all sides of image by a factor of three FWHM, and then truncated after convolution. In total, 3000 realizations of one-sample *t* statistic images were simulated for three different $n$, $n = 10$, 20, 30. Each *t* statistic image was based on $n$ realizations of smooth Gaussian random fields; to our knowledge there is no direct way to simulate smooth *t* fields. For each simulated dataset, a simple linear model was fit and residuals computed and used to estimate the smoothness, as in Ref. 27. To stress, for each realized dataset both the estimated and known smoothness was used for the random field inferences, allowing the assessment of this important source of variability.

For each simulated dataset we computed a permutation test with 100 resamples. While the exactness of the permutation test is given by exchangeability holding in these examples, this serves to validate our code and support other work.

We also simulated Gaussian statistic images with the same set of smoothness, but we did not apply the permutation test nor estimate smoothness.

For each realized statistic image we computed the Bonferroni, random field theory and permutation thresholds (except Gaussian) and noted the proportion of realizations for which maximal statistic was above the threshold, which is the Monte Carlo estimate of the familywise error in these null simulations.

For each realization we also computed three other FWER procedures: an FDR threshold, a threshold based on Gaussianizing the $t$ images, and a Bonferroni threshold using the estimated number of RESELS.

To estimate the equivalent number of independent test (see Section 2.5) we estimated $\theta$ with regression through the origin based on a transformation of equation (10):

$$\log(1 - F_{P_{(1)}}(t)) = \log(1 - t)V\theta \tag{32}$$

We replace $F_{P_{(1)}}(t)$ with the empirical cumulative distribution function of the minimum $P$-value found under simulation. Because we are generally interested in $\alpha_0 = 0.05$ we only use values of $t$ such that $0.03 \leq F_{P_{(1)}}(t) \leq 0.07$.

## 4.3 Simulation results

Figure 3 shows the accuracy in the smoothness estimate. As in Ref. 27, we found positive bias for low smoothness, although for higher smoothness we found slight negative bias. Positive bias, or overestimation of smoothness *under* estimates the degree of the multiple testing problem and can cause inferences to be anticonservative. (However, anticonservativeness was not a problem; see below.)

Figure 4 shows the results using the estimated smoothness. Figure 4(a) shows the permutation and true results tracking closely, while the RFT results are very conservative, only approaching truth for very high smoothness. Bonferroni is of course not

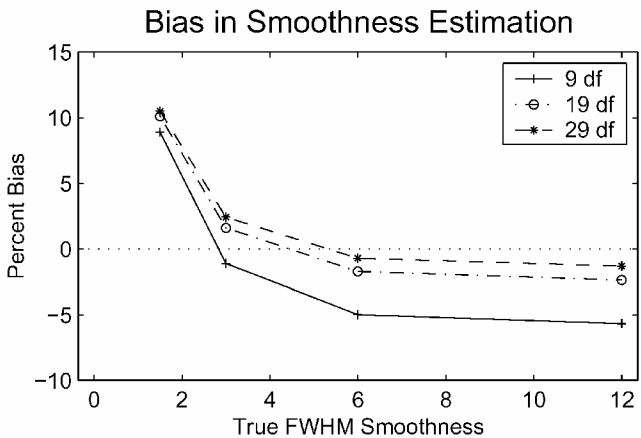


**Figure 3** Smoothness estimation bias as function of smoothness. Bias (estimated minus true) is smaller for larger DF and larger FWHM smoothness. Overestimated bias (for low smoothness) could result in anticonservative inferences (though apparently does not; see other results).

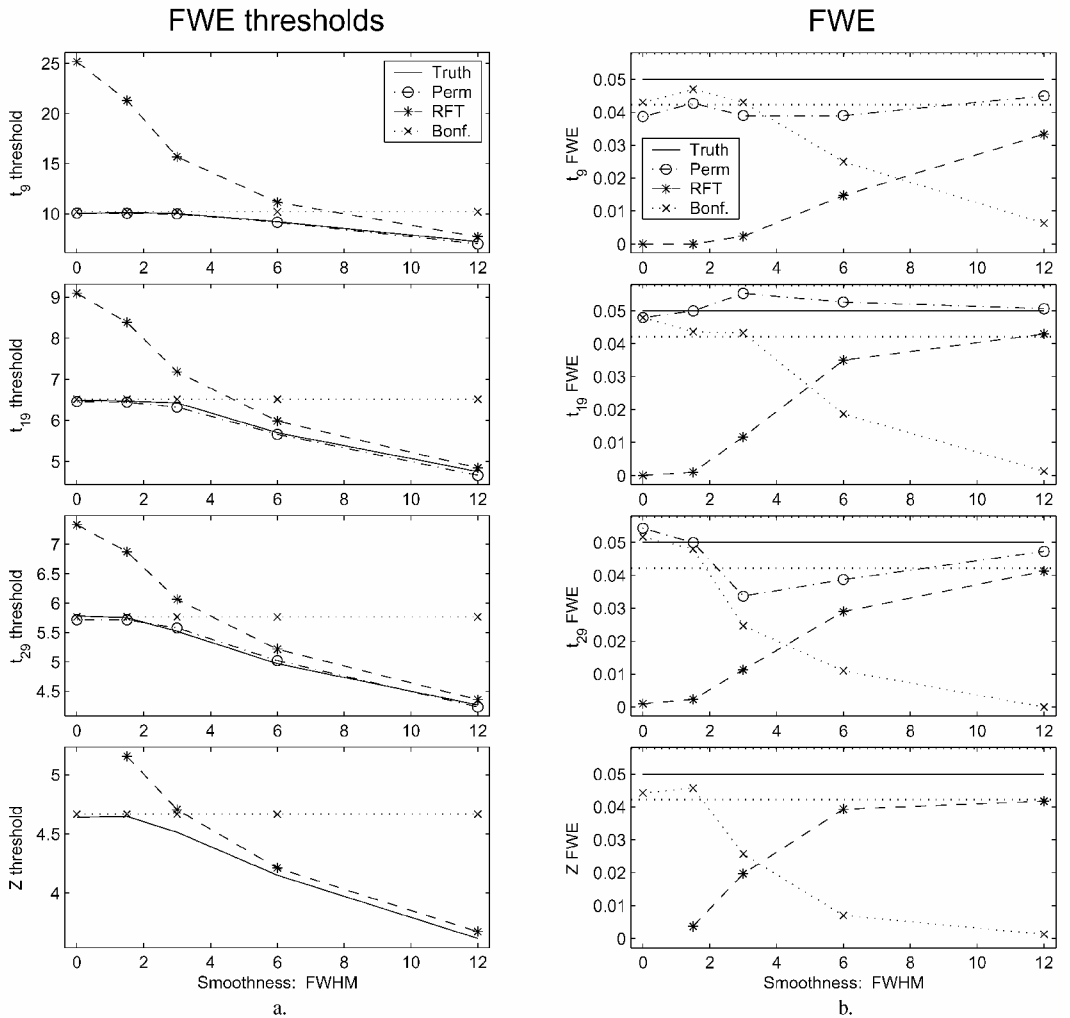


**Figure 4** Simulation results. (a) FWE threshold found by three different methods compared to truth. The Bonferroni threshold is nonadaptive, while permutation and random field methods both use lower thresholds with higher smoothness. The low-smoothness conservativeness of the random field thresholds intensifies with decreasing degrees of freedom. (b) Rejection rate of null simulations for a nominal $\alpha_0 = 0.05$ threshold, with a pointwise Monte Carlo 95% confidence interval shown with fine dotted line. The random field theory results are valid, but quite conservative for all but high smoothnesses. Bonferroni results are surprisingly satisfactory for up to three voxels FWHM smoothness, but then become conservative.

adaptive to smoothness, but is very close to truth for low smoothness, especially for low DF. The Gaussian results are much closer to truth than any of the $t$ results (note the $y$-axis range). Figure 4(b) shows the familywise error rates, which magnify performance differences. RFT is seen to be severely conservative for all but extremely smooth data, and Bonferroni is indistinguishable from truth for FWHM of three or less with DF of 9 and 19. The permutation performance is consistent with its exactness. For six FWHM and above, the Gaussian result is close to nominal.

Using true smoothness instead of estimated smoothness had little impact on the results. The rejection rates never differed by more than 0.003, except for the case of 9 DF and 12 FWHM, where it increased the rejection rate by 0.0084.

Figure 5 plots the cumulative density functions (CDFs) of the minimum *P*-value found by simulation, and compares it to other methods for approximating or bounding FWE. The CDF approximation provided by Bonferroni is the same for all figures, since the number of voxels is fixed. The RFT approximation (dash–dot line) changes with smoothness, but is far from true CDF for low smoothness and low DF; critically, for any given FWHM and DF, the RFT results do not improve with (decreasing *P*-value)

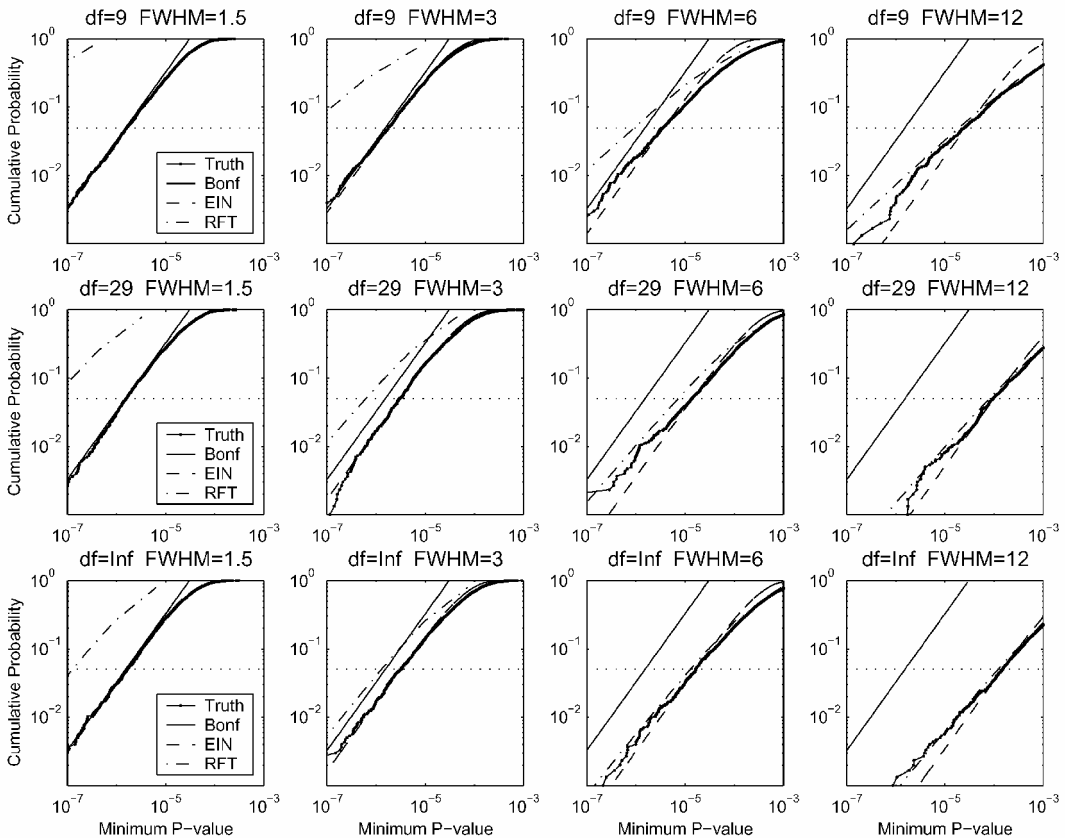


**Figure 5** Approximating minimum *P*-value distributions with FWE methods. The minimum *P*-value CDF obtained by simulation ('Truth', solid line with dots) is compared to three different approximations: Bonferroni inequality ('Bonf', solid line), random field theory ('RFT', dot–dashed line) and the equivalent independent *n* (EIN, dashed line); a corrected *P*-value of 0.05 is indicated (horizontal dotted line). These plots reflect the findings in Figure 4: Bonferroni is accurate for data as smooth as 1.5 FWHM data; RFT is more conservative than Bonferroni for data as smooth as three FWHM, and for six FWHM for 6 DF. While Figure 4 only depicted results for $\alpha_0 = 0.05$, note that for a given smoothness and DF the RFT results do not improve with more stringent thresholds (less than 0.05 corrected). For the 12 FWHM smoothness data the RFT results are quite accurate, and provide a better approximation than the equivalent independent *n* approach. Particularly for 9 DF and 12 FWHM, note that the EIN approach fails to have the correct slope (it intersects the true CDF around $F_{N_p}(0.05)$ by construction; see Section 2.5).

threshold. This indicates that the poor RFT performance is *not* due to use of an insufficiently high threshold. Finally, note that the CDF of an equivalent independent number (EIN) of observations (dashed line) follows the true CDF quite well for moderate smoothness, but at high smoothness it has the wrong slope and cannot match the CDF in general [as predicted by equations (30) and (31)]. That the EIN approach performs so well for moderate smoothness suggests that it may yet be a tenable theoretical approach.

For 9 DF point estimates for $\theta$ were found to be 0.90, 0.94, 0.87, 0.043 and 0.06 for 0, 1.5, 3, 6, 12 voxel FWHM smoothness, respectively. While Figure 5 indicates that the EIN approach is inappropriate for high smoothness, for three voxel FWHM smooth-

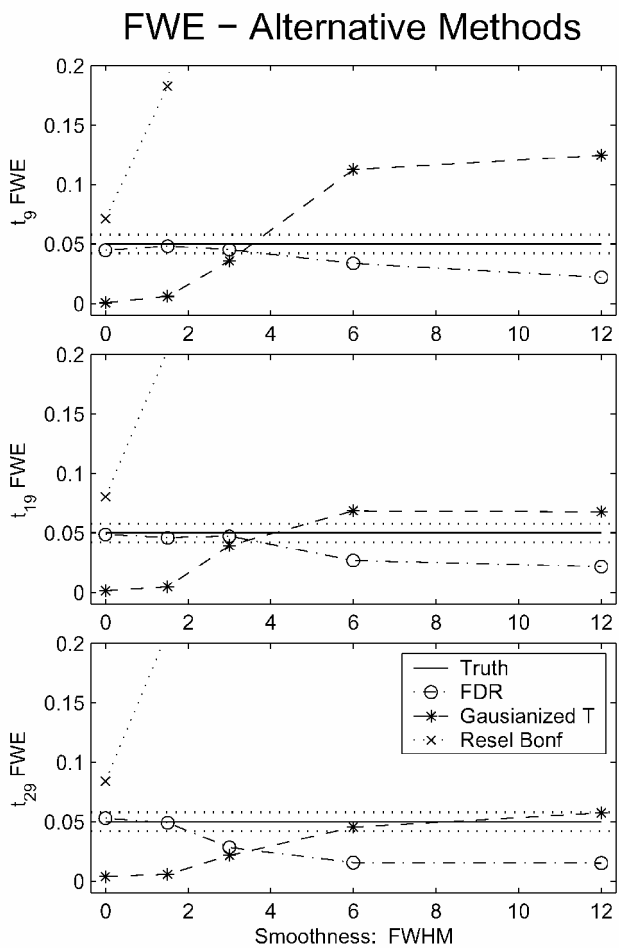


**Figure 6** Comparison of other FWE methods. The RESEL–Bonferroni approach fails to control FWE for any smoothness considered. The Gaussianized T approach does not reliably control FWE, in particular being anticonservative for smooth, low DF images. FDR does control FWE (weakly), but becomes somewhat conservative for increasing smoothness. Fine dotted line indicates pointwise Monte Carlo 95% confidence interval.

ness a $32^3$ voxel $t_9$ image has the same FWE threshold as $\theta V = 0.87 \times 32^3 = 28\,623$ independent voxels.

Figure 6 shows the performance of three alternative methods. The RESEL Bonferroni approach fails to control FWE, and for moderate to high smoothness exceeded a FWE of 0.5 (off the plot, not shown). The Gaussianized $t$ method exhibits conservativeness for low smoothness, but for low DF it is anticonservative, suggesting it would be inappropriate to use for all but the high DF. In this complete null simulation, Benjamini and Hochberg's FDR controls FWE, although it becomes somewhat conservative for increasing smoothness.

### 4.4 Results discussion

While some authors have observed RFT conservativeness,[36,50,51] other have not.[2,26] However, our findings are consistent with the literature, because the authors that found RFT results to be accurate used Gaussian data with high smoothness. For example, Worsley *et al.*[2] found the expected $\chi_u$ was quite accurate on $Z$ images, but the smoothness of their data was approximately 10 voxels FWHM. Our Gaussian simulations are consistent with this, and, for all but the lowest DF, our $t$ simulations also suggest that 10 FWHM is sufficient.

With our real data studies the permutation method was found to be more sensitive in all 11 datasets. This is consistent with our simulations, in particular that the RFT method was increasingly conservative for shrinking degrees of freedom. By conventional standards in functional neuroimaging our real data would be considered quite smooth (4–6 voxel FWHM), but our simulations indicate this is still insufficient for accurate RFT thresholds.

As a note on the selection of these datasets, they represent a three-year process of collecting group-level fMRI and PET datasets. The only data omitted were other effects from the studies included, usually other nonorthogonal contrasts with qualitatively similar results. In five years of applying these methods we have never seen a small DF dataset ($< 10$) where the $t$ random field method outperforms the permutation test.

## 5 Discussion

We have attempted to provide a comprehensive review and a representative comparison of FWE methods for functional neuroimaging. From Bonferroni and its extensions, to cutting-edge random field theory methods, to permutation methods of Fisher, we have attempted to cull all available tools that are relevant for the massive, dependent data of functional neuroimaging. With an assumption of positive dependence, we can make use of slightly improved Bonferroni methods. With an assumption of smoothness, we can make use of smoothness-adaptive RFT methods. And with few assumptions at all and some computational effort, we have both an adaptive and powerful method.

There are several limitations of these findings. First, yet more datasets should be studied, over yet a wider range of smoothnesses and group sizes. We have focused on very small group data to demonstrate a suspected conservativeness of RFT methods. However, more moderate group sizes are needed to see exactly when RFT methods lose power. Secondly, more simulations are needed for larger volumes, and for more

realistically shaped search regions. Our 32-cubed volume is too small when 1 $mm^3$ voxels are used and does not reflect the wrinkled-ellipsoidal topology of real brain data. And finally, the computational burden of the permutation tests must be considered, along with the flexibility of a general linear modeling tool combined with RFT inference.

## Acknowledgements

The authors would like to thank Keith Worsley for many valuable conversations on random field theory, especially Tor Wager for help with the pain data. The authors wish to thank all of the individuals who contributed data to our evaluations.

## References

1 Hochberg Y, Tamhane AC. *Multiple comparison procedures.* New York: Wiley, 1987.
2 Worsley K, Evans A, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism* 1992; **12**: 900–18.
3 Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ. Comparing functional (PET) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism* 1991; **11**: 690–99.
4 Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 2002; **64**: 479–498.
5 McShane LM. Statistical issues in the analysis of microarray data. In: *Proceedings of the International Biometrics Society*, Frieburg, Germany, July 2002.
6 Korn DL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data, Technical report. Biometric Research Branch, National Cancer Institute, Bethesda, Maryland 20892 USA, August 2001.
7 Westfall PH, Young SS. *Resampling-based multiple testing: examples and methods for p-value adjustment.* New York: Wiley, 1993.
8 Tong Y. *Probability inequalities in multivariate distributions.* New York: Academic Press, 1980.
9 Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**: 65–70.
10 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**: 800–802.
11 Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**: 751–54.
12 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, Methodological* 1995; **57**: 289–300.
13 Holland BS, Copenhaver MD. An improved sequentially rejective Bonferroni test procedure (Corr: V43 p. 737). *Biometrics* 1987; **43**: 417–23.
14 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**(4): 1165–88.
15 Sarkar SK. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics* 2002; **30**(1): 239–57.
16 Sarkar SK. Recent advances in multiple testing. Technical report. Philadelphia: Temple University, 2002.
17 Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 1995; **4**: 58–73.
18 Cao J, Worsley KJ. *Spatial statistics: methodological aspects and applications*, chapter 8. Applications of random fields in human brain mapping. Lecture Notes in Statistics; v. 159. New York: Springer, 2001; pp 169–82.
19 Petersson KM, Nichols TE, Poline J-B, Holmes AP. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philosophical Transactions of the Royal Society, Series B* 1999; **354**: 1261–81.

20 Adler RJ. *The geometry of random fields.* New York: Wiley, 1981.

21 Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism* 1992; **12**(6): 900–18.

22 Poline JB, Worsley KJ, Evans AC, Friston KJ. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage* 1997; **5**(2): 83–96.

23 Worsley KJ, Andermann M, Koulis T, MacDonald D, Evans AC. Detecting changes in nonisotropic images. *Human Brain Mapping* 1999; **8**: 98–101.

24 Poline J-B, Worsley KJ, Holmes AP, Frackowiak RSJ, Friston KJ. Estimating smoothness in statistical parametric maps: variability of *p* values. *Journal of Computer Assisted Tomography* 1995; **19**: 788–96.

25 Worsley KJ. Non-stationary FWHM and its effect on statistical inference of fMRI data. *Presented at the 8th International Conference on Functional Mapping of the Human Brain*, 2–6 June, 2002, Sendai, Japan. Available on CDRom. *NeuroImage* 2002; **16**(2): 779–80.

26 Holmes AP. *Statistical issues in functional brain mapping*, PhD thesis. University of Glasgow, Glasgow, 1994. Available from http://www.fil.ion.ucl.ac.uk/spm/papers/APH_thesis.

27 Kiebel S, Poline J, Friston K, Holmes A, Worsley K. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage* 1999; **10**: 756–66.

28 Worsley KJ. Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Annals of Statistics* 1995; **23**: 640–69.

29 Sampson PD, Guttorp P. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association* 1992; **87**: 108–19.

30 Hayasaka S, Nichols TE. A resel-based cluster size permutation test for non-stationary images. *Presented at the 8th International Conference on Functional Mapping of the Human Brain*, 2–6 June, 2002, Sendai, Japan. Available on CDRom. *NeuroImage* 2002; **16**(2): 1062–63.

31 Dinov ID, Mega MS, Thompson PM *et al.* Analyzing functional brain images in a probabilistic atlas: a validation of sub-volume thresholding. *Journal of Computer Aided Tomography* 2000; **24**(1): 128–38.

32 Good P. *Permutation tests. A practical guide to resampling methods for testing hypotheses.* New York: Springer Verlag, 1994.

33 Peasarin F. *Multivariate permutation tests: with applications in biostatistics.* New York: Wiley, 2002.

34 Efron B, Tibshirani R. *An introduction to the bootstrap.* Boca Raton: Chapman & Hall, 1993.

35 Davison AC, Hinkley DV. *Bootstrap methods and their application.* Cambridge: Cambridge University Press, 1997.

36 Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping* 2001; **15**: 1–25.

37 Scheffé H. Statistical inference in the non-parametric case. *Annals of Mathematical Statistics* 1947; **14**: 304–32.

38 Belmonte M, Yurgelun-Todd D. Permutation testing made practical for functional magnetic resonance image analysis. *IEEE Transactions on Medical Imaging* 2001; **20**: 243–48.

39 Liu C, Raz J, Turetsky B. An estimator and permutation test for single-trial fMRI data. In *Abstracts of ENAR Meeting of the International Biometric Society*, Pittsburgh, March 1998.

40 Brammer MJ, Bullmore ET, Simmons A *et al.* Generic brain activation mapping in functional mri: a nonparametric approach. *Magnetic Resonance Imaging* 1997; **15**: 763–70.

41 Locascio JJ, Jennings PJ, Moore CI, Corkin S. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping* 1997; **3**: 168–93.

42 Bullmore E, Long C, Suckling J. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Human Brain Mapping* 2001; **12**: 61–78.

43 Fadili J, Bullmore ET. Wavelet-generalised least squares: a new BLU estimator of regression models with long-memory errors. *NeuroImage* 2001; **15**: 217–32.

44 Holmes AP, Friston KJ. Generalisability, random effects & population inference. *NeuroImage* 1999; **7**(4) S754. *Proceedings of Fourth International Conference on Functional Mapping of the Human Brain,* 7–12 June, 1998, Montreal, Canada.

45 Henson RNA, Shallice T, Gorno-Tempini ML, Dolan RJ. Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex* 2002; **12**: 178–86.
46 Marshuetz C, Smith EE, Jonides J, DeGutis J, Chenevert TL. Order information in working memory: fMRI evidence for parietal and prefrontal mechanisms. *Journal of Cognitive Neuroscience* 2000; **12**(S2): 130–44.
47 Watson JDG, Myers R, Frackowiak RSJ *et al.* Area V5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex* 1993; **3**: 79–94.
48 Phan KL, Taylor SF, Welsh RC *et al.* Activation of the medial prefrontal cortex and extended amygdala by individual ratings of emotional arousal: An fMRI study. *Biological Psychiatry* 2003; **53**: 211–15.
49 Holmes AP, Blair RC, Watson JDG, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism* 1996; **16**(1): 7–22.
50 Stoeckl J, Poline J-B, Malandain G, Ayache N, Darcourt J. Smoothness and degrees of freedom restrictions when using spm99. *NeuroImage* 2001; **13**: S259.
51 Singh KD, Barnes GR, Hillebrand A. Group imaging of task-related changes in cortical synchronisation using non-parametric permutation testing. *NeuroImage* 2003; **19**: 1589–1601.