Localized Cluster Enhancement: TFCE Revisited with Valid Error Control

Samuel Davenport, Wouter Weeda, Thomas E. Nichols, Jelle J. Goeman

February 6, 2026

Abstract

Threshold-free Cluster Enhancement (TFCE) was introduced as a method to overcome the limitations of using cluster size inference for detecting effects in neuroimaging statistic maps. TFCE has seen wide application likely due to its high sensitivity, based on an image providing a voxel-wise representation of "cluster-like local spatial support". In this work we demonstrate that TFCE sensitivity comes at the cost of spatial specificity. Specifically, we show that TFCE does not provide voxel-wise or even cluster-wise error control. Most troubling, this means that an isolated TFCE detection can imply true signals almost anywhere in the brain. Embedding TFCE within a closed testing framework, we propose Localized Cluster Enhancement (LCE) which addresses these problems. We show that LCE can be used to detect activity within anatomical or data-driven regions of interest while controlling false positive rates.

1 Introduction

Threshold-free Cluster Enhancement (TFCE) (Smith and Nichols (2009), henceforth SN) is a widely used method for detection of effects in neuroimaging statistic maps, such as activations in fMRI or group differences with any modality. TFCE produces an image providing a voxel-wise representation of "cluster-like local spatial support", and was introduced with the aim of improving upon cluster size inference (Friston et al., 1994; Hayasaka and Nichols, 2003), in order to address problems of "smoothing, threshold dependence and localization", as indicated in the title of SN.

TFCE claims superiority over cluster size inference in two respects. The first claim is that TFCE is less dependent than cluster size inference on arbitrary, user-specified parameters, since it does not require specification of a threshold or an amount of smoothing. The second claim is that TFCE is better able to localize relevant effects than cluster size inference, providing "voxel-wise p-values" with "voxel-wise accuracy" (SN, p. 84).

In this paper, we will challenge both of these claims. In fact, we will show that the reverse of both claims is true. Regarding the arbitrariness of user-specified parameters, TFCE has more arbitrary

parameters than cluster size inference, with the original work using heuristics and simulations to justify fixed values for these parameters. One of these parameters explicitly plays the role of a threshold, and we show that this parameter is crucial for proper interpretation of the results of TFCE. Regarding localization, we refute the claim of voxel-wise accuracy, which was not formally proven by SN. In fact, we show that TFCE has weaker error control properties than cluster size inference. Where cluster size inference has cluster-wise error control, i.e., it infers with 95% confidence that at least one voxel in a significant cluster is active, TFCE only allows the weaker inference that an active voxel exists somewhere in the surroundings of a significant cluster. (As we are primarily motivated by fMRI, we will use the terms "active" and "activation" throughout, but we note that TFCE is also used for structural analyzes where group difference or covariate effects are of interest.) We show using numerical simulations that TFCE as classically used can thus have severely inflated voxelwise and clusterwise error rates.

In order improve the validity of TFCE we introduce a localized TFCE approach we call localized cluster enhancement (LCE). LCE arises from embedding TFCE within a closed testing framework (Marcus et al., 1976; Goeman and Solari, 2011). A similar embedding has recently been used to enhance the possibilities of cluster size inference (Goeman et al., 2023), allowing that method to make statements about the extent of activation with clusters. Similar extent-of-activation statements are difficult to derive for TFCE due to the way that the TFCE statistic integrates information. Still, the closed testing embedding allows us to establish several novel properties. In the first place, we precisely define the neighborhoods on which we can claim the existence of an active voxel. This results in a novel definition of a "TFCE-significant cluster". Secondly, we show how LCE, like cluster size inference (Goeman et al., 2023), can be used to infer the presence of activity in anatomical ROIs, even when these are defined after seeing the data. We will show formally that LCE controls cluster and regional error rates.

This paper is laid out as follows. In Section 2 we describe TFCE and analyze the claims that it makes about its error rate and being threshold free. In Section 2.3.3 we run numerical simulations which show that TFCE can have inflated voxelwise and clusterwise error rates. In Section 3 we establish results about the type of error control that is provided by TFCE. We introduce localized cluster enhancement which arises from embedding TFCE into a closed testing framework. In Section 5 we will study the performance of LCE in practice. Finally, in Section 6 we discuss our findings and provide advice to help ensure that future methods correctly control false positive rates.

Matlab code implementing LCE is available in the StatBrainz package (Davenport, 2024). In order to implement this we adapted code from the Matlab TFCE pacakge (`https://github.com/markallenthornton/MatlabTFCE`). A repository containing scripts to run all of the analyses performed in this paper is available at `https://github.com/sjdavenport/2024_clustersize_vs_tfce`. A MATLAB tutorial on localized TFCE is also available at `github.com/sjdavenport/StatBrainz/Tutorials/`.

2 Motivation

In this section we review the TFCE approach and its claims. We start with a complete description of the method and go on to examine its two primary claims: 1) TFCE is threshold free and 2) TFCE controls the voxelwise error rates.

2.1 Notation and description of TFCE

We start by revisiting TFCE and its claimed properties in more detail. We will follow the notation of SN as much as possible, but deviate from this as mathematical precision demands.

Let $\mathcal{B} \subset \mathbb{R}^3$ denote the set of voxels making up the brain. For every voxel $v \in \mathcal{B}$ suppose we have a test-statistic T_v, designed to test the voxel-specific null hypothesis that voxel v is not active. Let $\mathcal{E}(h) = \{v \in \mathcal{B} : T_v \geq h\}$, the excursion set, be the set of locations $v \in \mathcal{B}$, whose test-statistics T_v survive a threshold h. Cluster size inference is based on the size of the connected components in $\mathcal{E}(h)$ for a fixed value of h. That is, the size of each set of connected voxels within $\mathcal{E}(h)$ is tested against an empirical value k corresponding to the 95%-percentile of cluster sizes under the null hypothesis of no activation. TFCE, in contrast, integrates information for each set of connected voxels over a range of values of the threshold h. This integration over many thresholds motivates the name "threshold-free".

Given a connectivity criterion in $\mathbb{R}^3$, e.g., 26-connectivity, we can view the entire set of in-brain voxels $\mathcal{B}$ as a graph. Where $\mathcal{B}$ is generally fully connected (a path exists within the graph from any voxel to any other), the same does not hold for the thresholded set $\mathcal{E}(h)$. For each threshold h, we may split $\mathcal{E}(h)$ into its connected components, or clusters. For each $v \in \mathcal{E}(h)$, let $\mathcal{C}_v(h)$ be the connected component of $\mathcal{E}(h)$ which contains v, and for each $v \notin \mathcal{E}(h)$, let $\mathcal{C}_v(h) = \emptyset$, since v is not contained in any connected component. Define $e_v(h) = |\mathcal{C}_v(h)|$ to be the size (number of voxels) of this connected component. In other words, each $e_v(h)$ indicates the size of a cluster if v is within a supra-threshold cluster, or is zero when v is not in a cluster.

Based on $e_v(h)$, the TFCE statistic, for a voxel $v \in \mathcal{B}$ can be written as

$$S_v = \int_{h_0}^{T_v} h^H e_v(h)^E \mathrm{d}h. \tag{1}$$

Here, E, H and h_0 are tuning parameters, chosen by default as $E = 1/2$, $H = 2$ and $h_0 = 0$ (SN). E and H indicate the relative weight given to the size and the height of the cluster respectively. High values of E give more weight to clusters with a large spatial extent while high values of H give more weight to clusters which are taller. Essentially, the TFCE statistic of voxel v integrates information across all neighbouring supra-threshold voxels for each threshold from h_0 to $h = T_v$. Note that $e_v(h)$ thus changes for each threshold h. We illustrate this visually in Figure 1.

Though (1) is a test statistic for a specific voxel v, it is important to note that S_v does not only depend on the test statistic T_v of the voxel v itself, but also on the test statistics T_v, $v \in \mathcal{B}$ of other voxels (in practice S_v can depend on voxels from all over the brain, see Figure 2). This dependence is through

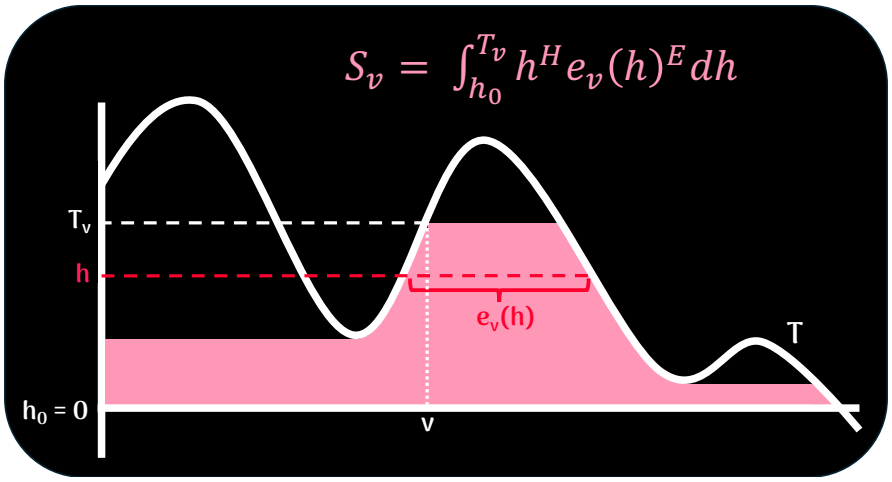

Figure 1: Illustrating the TFCE integral. The TFCE statistic S_v at a voxel v is the area of the shaded pink region. Note that as a result the value of S_v depends on the values of the original test-statistic in regions away from v.

$e_v(h)$, which depends on the entire supra-threshold set $\mathcal{E}(h)$. This fact will become important once we start to analyze the error control properties of TFCE in Section 3.

The integral in the definition of S_v is computationally intensive, so it is generally approximated by the following sum:

$$S_v \approx \delta \sum_{j=0}^{\lfloor (T_v - h_0)/\delta \rfloor} (h_0 + j\delta)^H e_v(h_0 + j\delta)^E. \tag{2}$$

The increments of this sum are specified by δ, which introduces a further tuning parameter that must be chosen by the user, and which is set at default at $\delta = 0.1$ in FSL (Smith et al., 2004).

The significance of S_v is determined using a permutation approach. Suppose that, for each $v \in \mathcal{B}$ X_v are the underlying data corresponding to voxel v, so that we can write $T_v = T(X_v)$. We make the dependence of S_v on $\mathbf{X} = (X_v)_{v\in\mathcal{B}}$ explicit by writing $S_v(\mathbf{X})$; this dependence arises due to the dependence of the extent $e_v(h)$ on the test-statistics $\{T_v\}_{v\in\mathcal{B}}$ meaning that in fact formally we have $e_v(h) = e_v(h, \mathbf{X})$. Let $\pi_1, \ldots, \pi_P$ be appropriately chosen permutations, and apply these to $\mathbf{X}$ to get permuted data $\mathbf{X} \circ \pi_i = (X_v \circ \pi_i)_{v\in\mathcal{B}}$, and resulting TFCE statistics $S_v(\mathbf{X} \circ \pi_i)$ for every $i = 1, \ldots, P$ and $v \in \mathcal{B}$. Then, given $\alpha \in (0, 1)$, a critical value t^* for the TFCE statistic can now be calculated as the $(1 - \alpha)$-quantile of the distribution of $\max_{v\in\mathcal{B}} S_v(\mathbf{X} \circ \pi_i)$ over the permutations. TFCE claims as significant all voxels v for which $S_v > t^*$. We will call clusters of such voxels, namely the connected components of $\mathcal{E}^* = \{v \in \mathcal{B} : S_v \geq t^*\}$, 'TFCE-significant clusters'.

Remark 1. *As a concrete of this framework we could for instance take X_v to be the concatenated vector of observations from two groups of subjects at the voxel v and T_v to be the two-sample test-statistic. Permutations then arise by shuffling the data between groups (jointly over all voxels).*

2.2 First claim: Threshold-free

TFCE - threshold free cluster enhancement - claims very directly to be threshold free. SN criticize the need of cluster size inference to define a cluster-forming threshold as an arbitrary choice about which no clear objective advice is available. Indeed, cluster size inference requires an a priori chosen threshold, which is arbitrary, but conventionally set to 3.1 (Goeman et al., 2023). By avoiding such a threshold TFCE claims to 'avoid large changes in the output being caused by small changes in the input' and 'to remove the dependence on the arbitrary choice of the cluster-forming threshold'.

However, in (1) we see that, while removing the classical cluster-forming threshold, TFCE adds three new tuning parameters: E – the exponent of the cluster extent, H – the exponent of the height and h_0 – the lower bound of the integration. (And, to be complete, the numerical approximation (2) requires even a fourth parameter δ.) Thus, despite the goal of SN to limit the number of free parameters, this number is in fact increased. The default values for $E = 0.5$ and $H = 2$ are empirically motivated, but it is difficult to call these choices 'objective'. And while a formal calculation (SN, Appendix C), relating TFCE to Fisher's combining method and random field theory, suggests $E = 2/3$, $E = 1/2$ is used instead.

The tuning parameter h_0, the lower bound of the integral in the calculation of TFCE, is of particular interest in the context of this paper as it acts as a threshold. Unfortunately, its default value of $h_0 = 0$ is not discussed at all in the original paper, despite how (as we shall see in Section 5) its choice can have a big impact in practice. The reason that h_0 acts as a threshold is that the support at the level h_0 determines the rejection regions over which TFCE has valid error control - just as the cluster defining threshold does for cluster size inference. We demonstrate this formally in Section 4.2, showing that TFCE is not in fact threshold free.

2.3 Second Claim: Error control

2.3.1 Error rates in neuroimaging

For neuroimaging based inference, it is important to make a distinction between voxel-wise or cluster-wise methods. When a voxel-wise method finds a significant cluster, it makes the claim that all voxels in the significant cluster are active. A cluster-wise method, in contrast, only makes the much weaker claim that a significant cluster contains at least one active voxel. The distinction between voxel-wise and cluster-wise make a big difference for the interpretation of results. Whereas a voxel-wise method is able to pinpoint the activation precisely, a cluster-wise method only returns a region in which activation is present—somewhere. Cluster-wise inference has been criticized as paradoxical (Woo et al., 2014; Rosenblatt et al., 2018). For example, with cluster-wise inference, unlike with voxel-wise, a large cluster is less informative than a small cluster, because the signal found is less well localized.

Proper error control should be tailored to the type of statements a method makes. Cluster-wise familywise error control at level α requires that with $1 - \alpha$ confidence all significant clusters contain at least one active voxel. Voxel-wise familywise error control means that with $1 - \alpha$ confidence all voxels contained within them are active. Cluster size inference has proven cluster-wise error control [REF]

[SD: Do we have a good reference for this? Other than Jelle's paper that is.]. Recently (Goeman et al., 2023) has shown that it is possible to make slightly stronger statements about the significant clusters found using cluster size inference however method does not make voxel-wise statements.

2.3.2 Examining the error control claimed by TFCE

In this subsection we take a closer look at the claims made by SN and in the literature regarding the error control provided by TFCE. In particular, Smith and Nichols (2009) claim that TFCE provides the stronger voxel-wise control, saying that TFCE returns "voxel-wise p-values [...] corrected for multiple comparisons", and is subsequently described as having "strong control over familywise error." The voxelwise interpretation has also made its way into the literature. For instance Li et al. (2017, p. 1270) say "For inference, voxel-level permutation testing is used to turn the TFCE image into voxel-wise P-values". Moreover on the FSL website `https://web.mit.edu/fsl_v5.0.10/fsl/doc/wiki/Randomise(2f)UserGuide.html` the guidance indicates that "TFCE [...] is a new method for finding "clusters" in your data without having to define clusters in a binary way. Cluster-like structures are enhanced but the image remains fundamentally voxelwise".

However, some authors have cast doubt on the voxel-wise error control property of TFCE. Spisák et al. (2019) present a variant of TFCE that, as they claim, provides "stricter control" of false positive voxels, implying that control of the original TFCE was inadequate. See for example Woo et al. (2014, p. 413): "However, TFCE is also subject to the same limitations of low spatial specificity when significant clusters are large".

SN do not provide a formal proof of voxel-wise control and instead rely on the informal argument that they use "standard permutation testing" (p. 84), which is later (p. 85) described in detail, and can be recognized as the single-step max-T method of Westfall and Young (1993). However, this method relies on the crucial assumption, not explicily checked by SN, that the joint distribution of the test statistics of inactive voxels is invariant under permutation (Westfall and Young, 1993; Goeman and Solari, 2010). In fact, this assumption is violated when test statistics of inactive voxels may depend on data from active voxels, as is the case for the TFCE statistic (1) through the term $e_v(h)$. The type of error control provided by TFCE is thus unclear from the original paper. As part of this work we examine the TFCE error rates empirically in Section 2.3.3 and fully characterize them theoretically in Section 3.

2.3.3 Empirical validation of TFCE error rates

In this subsection we empircally test whether TFCE in fact controls voxel and cluster-wise error rates. We consider the voxel and cluster level error of TFCE in simulations and compare to cluster size inference. To do so we run simulations in which we take a fixed signal, add noise to create realistic observations and generate realisations of this process. We consider a range of different simulation settings in which we vary the number of true clusters and their size, Figures 6 and S1 for an illustration of the signal structure. For noise we smooth Gaussian white noise with an isotropic Gaussian kernel with three different levels

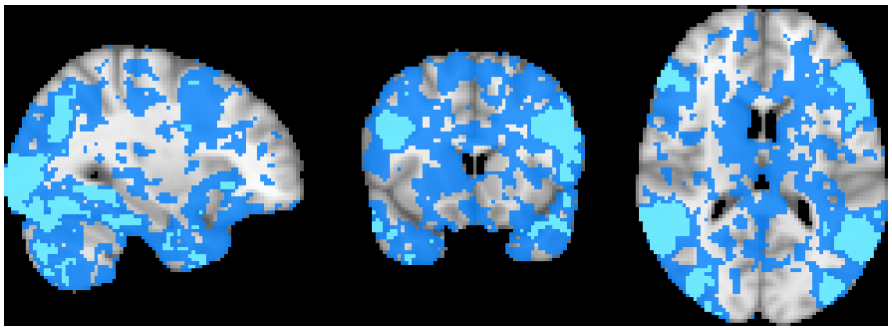

Figure 2: The true support of TFCE clusters.*[SD: Could add other brain examples here.]*

of applied smoothness (FWHM = $2, 4, 6$ voxels). In each simulation setting we generate $n_{\text{sim}} = 1000$ realisations.

For each simulation we run TFCE (with the default parameters i.e. $h_0 = 0$, $E = 0.5$, $H = 2$ and $\delta = 0.1$), compute the TFCE transformed test-statistic S_v at each $v \in \mathcal{B}$. We then use permutation to determine the TFCE threshold t^* and TFCE significant clusters, as described in Section 2.1. Let $\mathcal{N} \subseteq \mathcal{B}$ denote the subset of $\mathcal{B}$ on which there is no signal. Then in any given simulation, a voxelwise error occurs if $S_v > t^*$ for some $v \in \mathcal{N}$. Instead a cluster size error occurs if there is some TFCE significant cluster which lies entirely within $\mathcal{N}$. Aggregating over all simulations we can compute voxel and clusterwise error rates, see Section S1 for further details. We also compare against the errors obtained by cluster size inference, with the default cluster defining threshold of 3.1, in the same simulation settings.

The results are shown in Figure 3, XXX and XXX. They show that TFCE can have both inflated cluster and voxelwise error rates. Clusterwise inference has inflated voxelwise error rates (as it is not designed to control them) but crucially correctly controls the cluster error rate, as proven in Goeman et al. (2023). Results for other simulated settings were similar, see Appendix XXX for further details.

From these simulations we conclude that TFCE does not guarantee error control either at the voxel level or at the cluster level. So what error rate does it control? We provide answers to this in the following section in which we establish an improved TFCE based approach which has provably guaranteed error control.

3 Theory

The theory in this section will apply not just to TFCE, but for a class of methods of which TFCE is a special case. In particular we will work with a generalized form of the test statistic (1). For functions f and g, define

$$S_v = \int_{h_0}^{\infty} f(h) g\big(e_v(h)\big) \mathrm{d}h. \tag{3}$$

This reduces to the TFCE-statistic (1) if we take $f(x) = x^H$ and $g(x) = x^E$, and note that $e_v(h) = 0$ if $h > T_v$. However, the general formula (1) not only describes TFCE, but also encompasses other

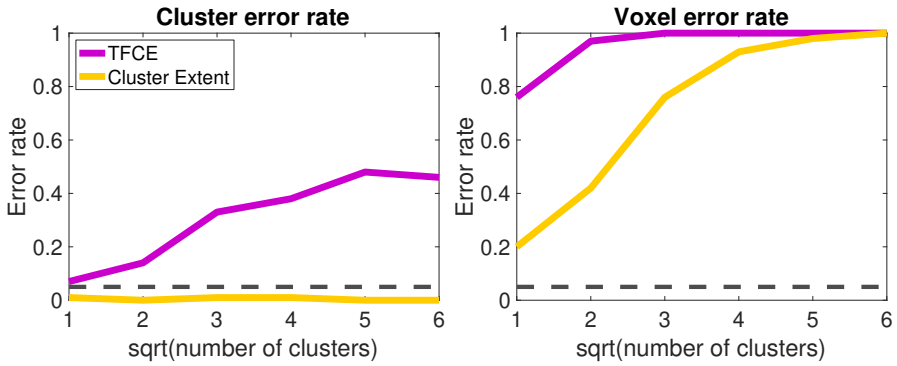

Figure 3: Clusterwise (left) and voxelwise (right) error rate for TFCE versus cluster size inference in simulations. TFCE has inflated cluster error rates whilst cluster size inference correctly controls them to a level $\alpha = 0.05$. Both methods have inflated voxelwise error rates. *[SD: Consider adding clusterwise and voxelwise LCE lines.]*

methods for other choices of f and g. We recover cluster size inference by taking f to be a Dirac δ-function shifted to h_0, and g the identity; cluster-mass inference with $f(x) = 1$ and g the identity; and peak height inference with $f(x) = 1$ and $g(x)$ the indicator of $x > 0$. Probabilistic TFCE can also be obtained, when f and g are reconstructed according to Spisák et al. (2019, pp. 14 and 17). In what follows we will not make many assumptions on f, g and h_0, except that g must be non-decreasing, with $g(0) = 0$, and $f(h)$ must be non-negative for $h \geq h_0$. This ensures that increases in the extent and height of the test-statistic lead to higher values of S_v. We also assume that f and g are fixed, and not data-dependent. However, as discussed in Section 2.1 for the TFCE statistic (1), S_v is a function of the data and we can make this explicit by writing $S_v = S_v(\mathbf{X})$.

3.1 Permutation Assumptions

Since TFCE uses permutation-based inference we must make the standard assumptions, which are typically required by permutation-based multiple testing methods, in order to ensure valid inference. Let $\mathcal{I} \subseteq \mathcal{B}$ be the collection of all inactive voxels, and Π be the group of permutations considered. We assume that the distribution of the data for this inactive set is invariant to the permutations we use. This is stated formally in Assumption 1.

Assumption 1. For every $\pi \in \Pi$, we have $(\mathbf{X}_v)_{v\in\mathcal{I}} \overset{d}{=} (\mathbf{X}_v \circ \pi)_{v\in\mathcal{I}}$.

This assumption is for instance satisfied for one-sample testing when sign-flipping to flip the contributions, under the assumption of symmetry. It is also satisfied for two-sample group testing, with the permutations involving mixing the groups, under the null hypothesis that the distribution of the data in each group has the same distribution (Nichols and Holmes, 2002).

3.2 TFCE provides weak familywise error rate control

3.2.1 Constructing a global test

We will first show that TFCE, in this general form (3), can be used to construct a valid hypothesis test for $H_{\mathcal{B}}\colon \mathcal{I} = \mathcal{B}$, the global null hypothesis that all voxels in the brain are inactive. Given an error rate $\alpha \in (0, 1)$, this test rejects $H_{\mathcal{B}}$ if there exists a voxel for which the TFCE-statistic is above the threshold t^*. This threshold is defined as the kth order statistic of

$$\max_{v\in\mathcal{B}} S_v(\mathbf{X}\circ\pi_1), \ldots, \max_{v\in\mathcal{B}} S_v(\mathbf{X}\circ\pi_P),$$

where $k = \lceil(1-\alpha)P\rceil$. *[SD: Maybe we need to change this to be a bit clearer about what t^* is. JG: maybe just say that it is the kth smallest value among these, or the kth in line after sorting from small to large?]*

Theorem 2. *Suppose that Assumption 1 holds. Suppose that π_1 is the identity permutation and $\pi_2, \ldots, \pi_P$ are drawn independent and uniformly, with replacement, from Π. Then, if $H_{\mathcal{B}}$ is true,*

$$\mathrm{P}\big(\max_{v\in\mathcal{B}} S_v(\mathbf{X}) > t^*\big) \leq \alpha.$$

Theorem 2 asserts weak familywise error control for TFCE. It shows that, if $H_{\mathcal{B}}$ is true and there are no active voxels in the brain whatsoever, then TFCE will not give any significant result with probability at least $1-\alpha$. This means that the TFCE statistics can be used as a global test to detect the presence of any signal. Weak FWER control is a prerequisite for both voxelwise and clusterwise control, but does not imply either of those error rates.

3.2.2 Constructing a local test

The global test just constructed can be turned into a local test. Where a global test detects the presence of any signal in the entire brain, local tests detect the presence of any signal in a subset of the brain. A local test for a subset $\mathcal{R} \subseteq \mathcal{B}$, of the brain, tests the local null hypothesis $H_{\mathcal{R}}\colon \mathcal{R} \subseteq \mathcal{I}$ that all voxels in $\mathcal{R}$ are inactive.

To construct a local test for a subset $\mathcal{R}$, we can simply apply TFCE to the masked data $\mathbf{X}_{\mathcal{R}} = (X_v 1[v \in \mathcal{R}])_{v\in\mathcal{B}}$ in which all voxels in $\mathcal{B} \setminus \mathcal{R}$ are masked out: here $1[\cdot]$ denotes the indicator function. After masking, we can simply apply Theorem 2, letting $\mathcal{R}$ take the role of $\mathcal{B}$. It is worth spelling out the result. Define the critical value $t^*_{\mathcal{R}}$ as the kth order statistic of

$$\max_{v\in\mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}\circ\pi_1), \ldots, \max_{v\in\mathcal{B}} S_v(\mathbf{X}_{\mathcal{R}}\circ\pi_P),$$

where $\mathbf{X}_{\mathcal{R}}\circ\pi = (X_v\circ\pi)_{v\in\mathcal{R}}$ and $k = \lceil(1-\alpha)P\rceil$. Then we have the following version of Theorem 2 within the region $\mathcal{R}$.

Theorem 3. *Suppose that Assumption 1 holds. Suppose that π_1 is the identity permutation and $\pi_2, \ldots, \pi_P$ are drawn independent and uniformly, with replacement, from Π. Then, if $H_{\mathcal{R}}$ is true,*

$$\mathrm{P}\big(\max_{v\in\mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) > t^*_{\mathcal{R}}\big) \leq \alpha.$$

Theorem 3 implies that we may reject some $H_{\mathcal{R}}$, asserting that there is signal present in the subset $\mathcal{R}$ of the brain when TFCE, applied to a restricted data set in which all other voxels have been masked out, contains at least on significant voxel. Here, it is crucial that the TFCE statistics are recalculated after the restriction. We cannot simply take the maximum over $\mathcal{R}$ of the usual TFCE statistic: $\max_{v\in\mathcal{R}} S_v(\mathbf{X})$ as this is generally larger than $\max_{v\in\mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}})$, and the result of Theorem 3 ceases to apply if the latter is replaced by the former.

4 Localized Cluster Enhancement

In this section we will rebuild the TFCE method and derive an alternative definition of a TFCE-significant cluster, for which we can prove clusterwise error control. We will do this by embedding TFCE into a closed testing procedure (Marcus et al., 1976; Genovese and Wasserman, 2006; Goeman and Solari, 2011) in much the same way that cluster size inference was embedded into a closed testing procedure by Goeman et al. (2023), enhancing the properties of that method. Note that Goeman et al. (2021) showed that all methods controlling either voxelwise or clusterwise error can be uniformly improved by such an embedding, if they are not already equivalent to a closed testing procedure.

[SD: Write a TLDR to readers here saying summarizing what we do, and why and suggesting that they skip ahead to section 5 for the results etc.]

The statement of Theorem 3 is not yet corrected for multiple testing. The result is, therefore, useful only if $\mathcal{R}$ is a single region of interest, chosen independently of the data. In practice, we wish to infer on many such regions, and they are often chosen in a data-driven way. To handle this situation, we need clusterwise familywise error control over all regions $\mathcal{R}$ called significant. In order to achieve this we introduce Localized Cluster Enhancement - a procedure which arises as a closed testing embedding of TFCE (Goeman et al., 2021). As we shall show, LCE provides valid error control (simultaneously over all considered regions) while still enhancing the power to detect effects of interest, as shown in Section 5.

4.1 Simultaneous Error Control

We introduce Localized Cluster Enhancement (LCE) as follows. Given a region $\mathcal{R} \subseteq \mathcal{B}$, we can calculate an LCE-corrected p-value as

$$\text{LCE}(\mathcal{R}) = \frac{1}{P}\sum_{p=1}^{P} 1\left[\max_{v\in\mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) \leq \max_{v\in\mathcal{B}} S_v(\mathbf{X}\circ\pi_p)\right]. \tag{4}$$

It is the proportion of permutations for which the largest permutation brain-wise LCE statistic is larger than the largest region-wise LCE statistic in the real data. Then, given a level $\alpha \in (0,1)$, we reject the null hypothesis $H_{\mathcal{R}}$, for a given region $\mathcal{R}$, if $\text{LCE}(\mathcal{R}) \leq \alpha$. This procedure has valid error control simultaneously over all considered regions, providing valid adjusted p-values, as established in the following theorem.

Theorem 4. *(Simultaneous validity of LCE) Suppose that Assumption 1 holds. Suppose that π_1 is the identity permutation and $\pi_2, \ldots, \pi_P$ are drawn independent and uniformly, with replacement, from Π. Given $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(\textit{LCE}(\mathcal{R}) \geq \alpha \text{ for all } \mathcal{R} \subseteq \mathcal{I}\right) \geq 1 - \alpha.$$

Equivalently $\mathbb{P}\big(\text{there exists } \mathcal{R} \subseteq \mathcal{B} \text{ such that } H_{\mathcal{R}} \text{ is true and } \textit{LCE}(\mathcal{R}) \leq \alpha\big) \leq \alpha$.

This guarantee is valid for all $\mathcal{R} \subseteq \mathcal{I}$. As such it is also valid for data driven regions which can be selected by the user based on the data. With this guarantee we can be $(1-\alpha)100\%$ sure that any rejected region is in fact not null. Moreover we have the following characterization of the LCE test procedure.

Corollary 1. *Let t^* be the quantile (**??**) of the TFCE test-statistics. Then LCE rejects a region $\mathcal{R} \subseteq \mathcal{B}$ if and only if $\max_{v \in \mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) > t^*$. As such, under the assumptions of Theorem 4,*

$$\mathbb{P}\big(\text{there exists } \mathcal{R} \subseteq \mathcal{B} \text{ such that } H_{\mathcal{R}} \text{ is true and } \max_{v \in \mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) > t^*\big) \leq \alpha.$$

This result allows LCE to be used to identify significant regions with minimal extra computation cost relative to TFCE, whilst being able to make much stronger theoretical statements. The restriction of the calculated TFCE statistics, via the masking procedure (c.f. Figure 5), is what allows the inference to be interpreted locally instead of globally as in Section 3.2.1.

4.2 Clusterwise error control

The validity of the LCE adjusted p-values for any chosen region $\mathcal{R}$ means that these regions can be chosen to be data-driven. In particular, as formalized in Corollary **??** below, we can apply LCE to the TFCE significant clusters to determine whether they are in fact significant. Note that, as shown in Section XXX, TFCE does not control clusterwise error rates and so cannot be used to determine cluster significance of the TFCE selected clusters without the LCE enhancement.

Theorem 5. *(Establishing the significance of clusters identified by TFCE) Under the Assumptions of Theorem 4, let $\mathcal{C}_1, \ldots, \mathcal{C}_m \subseteq \mathcal{B}$ denote the clusters identified to be "TFCE significant". Then*

$$\mathbb{P}\left(\textit{LCE}(\mathcal{C}_i) \geq \alpha \text{ for all } 1 \leq i \leq n \text{ such that } \mathcal{C}_i \subseteq \mathcal{I}\right) \geq 1 - \alpha.$$

An example of doing so is shown in Figure XXX. In this example, applying LCE to the TFCE signficant clusters correctly identifies that one of these clusters is just due to noise, allows a correction of TFCE. Indeed, we can see from Figure XXX and Figures XXX and XXX, that LCE applied to the TFCE clusters provides valid clusterwise error control. We are finally in a position to characterize the actual error control provided by the original TFCE procedure. In particular TFCE as used in practice, provides error control with respect to the support of the test-statistic. To show this, for each threshold $h_0 \in \mathbb{R}$, given a connected component $\mathcal{D}$ of the set $\{v \in \mathcal{B} : T_v \geq h_0\}$, and a region $\mathcal{R} \subseteq \mathcal{D}$ define

$$\text{supp}_{h_0}(\mathcal{R}) := \mathcal{D}. \tag{5}$$

We can then formalize the error control of TFCE as follows.

Algorithm 1 Localized Cluster Enhancement

Require: Data $\mathbf{X} = (X_v)_{v\in\mathcal{B}}$, real-valued functions f, g, a threshold $h_0 \in \mathbb{R}$, a set of n regions of interest $\{\mathcal{R}_i\}_{i=1}^n \subseteq \mathcal{B}$, a number of permutations P, and a collection of permutations Π.

Compute the original test-statistics $\{T_v\}_{v\in\mathcal{B}}$.

Compute the transformed statistics $S_v(\mathbf{X}) = \int_{h_0}^{\infty} f(h)g\big(e_v(h, \mathbf{X})\big)\mathrm{d}h$, for each $v \in \mathcal{B}$.

Draw permutations $\pi_1, \ldots, \pi_P$ independently from Π.

for $p = 1$ to P **do**

    Compute the permuted transformed map: $S_v(\mathbf{X} \circ \pi_p) = \int_{h_0}^{\infty} f(h)g\big(e_v(h, \mathbf{X} \circ \pi_p)\big)\mathrm{d}h$, for $v \in \mathcal{B}$.

end for

for $i = 1$ to n **do**

    Mask the data to the region $\mathcal{R}_i$, to compute $\mathbf{X}_{\mathcal{R}_i} = (X_v 1[v \in \mathcal{R}_i])_{v\in\mathcal{B}}$.

    Compute the masked transformed map $S_v(\mathbf{X}_{\mathcal{R}_i}) = \int_{h_0}^{\infty} f(h)g\big(e_v(h, \mathbf{X}_{\mathcal{R}_i})\big)\mathrm{d}h$.

    Calculate $\mathrm{LCE}(\mathcal{R}_i) = \frac{1}{P}\sum_{p=1}^{P} 1\big[\max_{v\in\mathcal{R}_i} S_v(\mathbf{X}_{\mathcal{R}_i}) \leq \max_{v\in\mathcal{B}} S_v(\mathbf{X} \circ \pi_p)\big]$.

end for

return $\{\mathrm{LCE}(\mathcal{R}_i)\}_{i=1}^n$: the set of LCE-adjusted p-values.

given $\alpha \in (0, 1)$, reject $H_{\mathcal{R}_i}$ for each $1 \leq i \leq n$ such that $\mathrm{LCE}(\mathcal{R}_i) \leq \alpha$.

Figure 4: Pseudocode for Localized Cluster Enhancement (LCE).

Corollary 2. *Under the Assumptions of Theorem 4, let $\mathcal{C}_1, \ldots, \mathcal{C}_m \subseteq \mathcal{B}$ denote the clusters identified to be "TFCE significant". Then $\mathbb{P}\left(\mathit{supp}_{h_0}(\mathcal{C}_i) \subseteq \mathcal{I} \text{ for some } 1 \leq i \leq m\right) \leq \alpha$.*

Note that the statement of this corollary is well-defined since every TFCE significant cluster is a subset of a connected component of $\{v \in \mathcal{B} : T_v \geq h_0\}$ by construction. Corollary 2 shows that TFCE, as widely used in practice, is in fact able to make the statement that there is some voxel active within the support of the cluster - though not the cluster itself. As shown in Figure 2, this support can be very large (potentially covering a large part of the brain) and so this statement is still rather weak in practice. This result also illustrates the (strong) dependence of TFCE on the threshold h_0.

4.3 Voxelwise error control

Remarkably, it is also possible to make certain voxelwise statements using LCE. To see this note that for each $v \in \mathcal{B}$, taking $\mathcal{R} = \{v\}$, the restricted TFCE statistic is simply

$$S_v(\mathbf{X}_{\{v\}}) = \int_{h_0}^{T_v} f(h)g(1)\, dh, \tag{6}$$

since $e_v(h) = 1[h_0 \leq h \leq T_v]$. In the default setting where $f(h) = h^H$ and $g(1) = 1$, (6) reduces to

$$S_v(\mathbf{X}_{\{v\}}) = \int_{h_0}^{T_v} h^H\, dh = \frac{1}{H+1}(T_v^{H+1} - h_0^{H+1}),$$

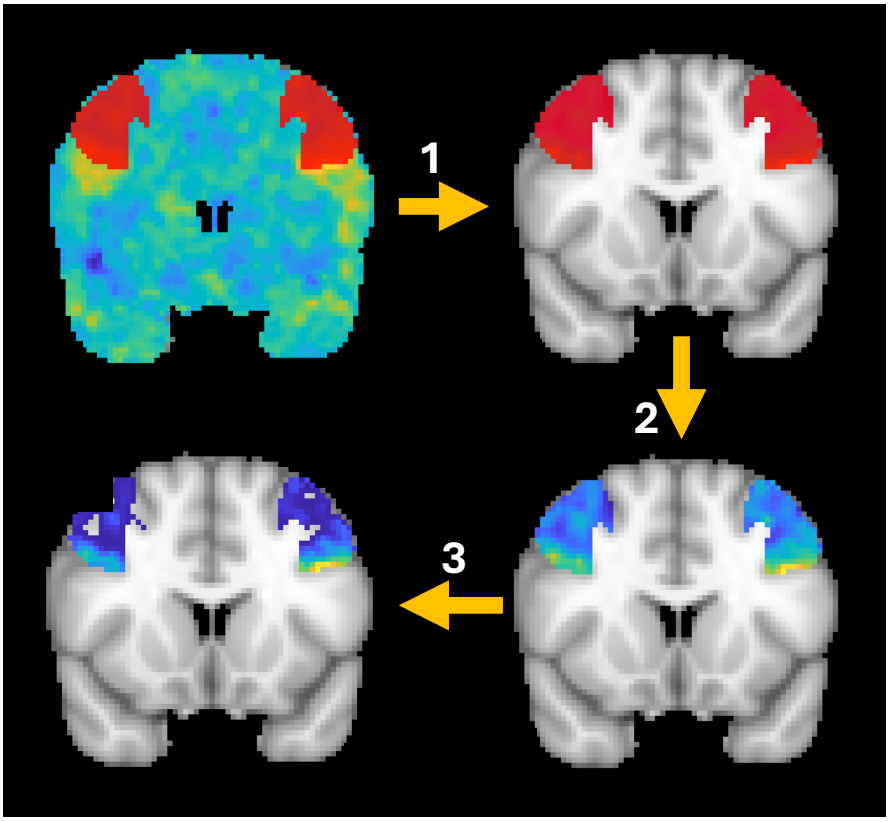

Figure 5: Localized Cluster Enhancement for a region $\mathcal{R}$. *[SD: Let's discuss how to make the most of this figure.]*

Applying Corollary 1 we can say then say that all voxels $v \in \mathcal{B}$ for which $T_v \geq (t^*(H+1)+h_0^{H+1})^{\frac{1}{H+1}}$, are voxelwise significant, whilst providing strongly control of the familywise error rate. We formalize this result in Section **??**. Under the default setting, in which $H = 2$, we have $\frac{1}{H+1}(T_v^{H+1} - h_0^{H+1})$, which is $\frac{T_v^3}{3}$ if $h_0 = 0$. As such, we may call all voxels such that $T_v > (3t^*)^{1/3}$ voxelwise significant and provide strong control of the FWER.

5 Results

5.1 Cluster error control using LCE

Include Figures with the cluster error of LCE for the threshold of 0 and 3.1. As well as for some notion of regional error.

Remark 6. *Note that when used for clusterwise inference, as discussed in Section, LCE like cluster-size inference and TFCE does not control the voxel-wise false positive rates. We illustrate this in Section*

SXXX. A form of voxelwise error control is possible, however, using LCE as shown in Section XXX.

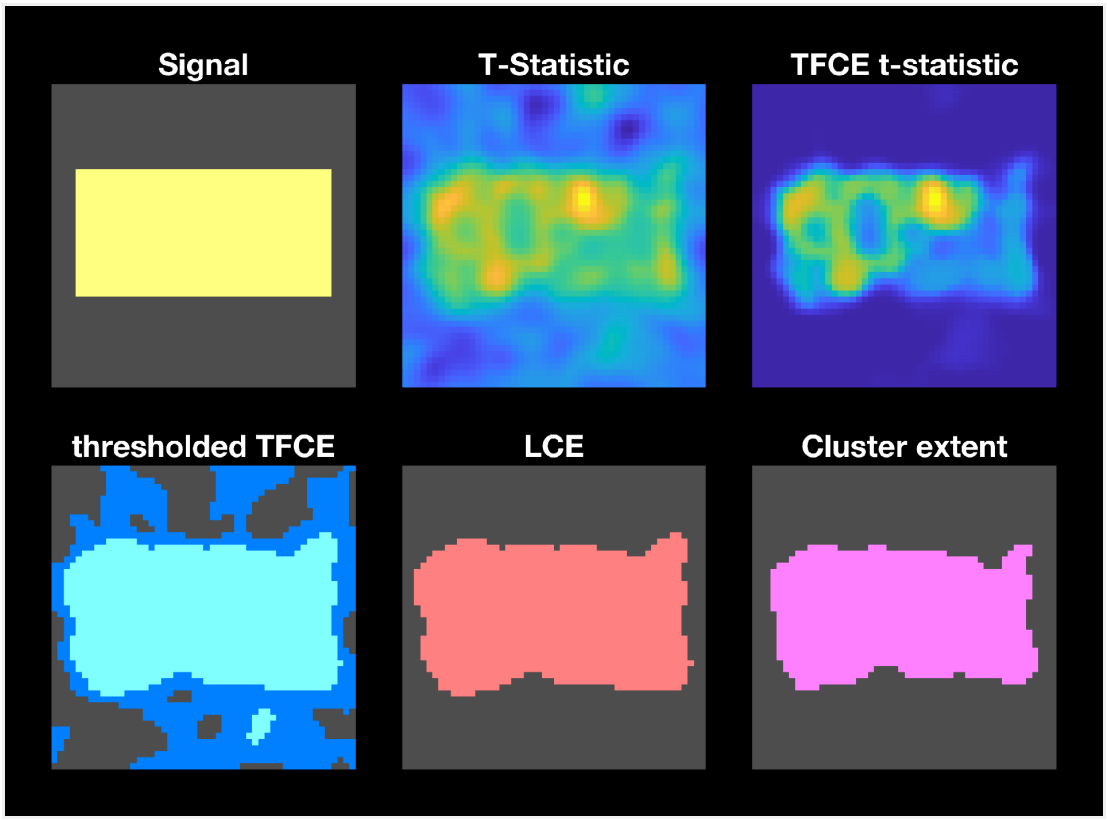

Figure 6: Comparing TFCE, LCE and cluster size inference in a simple simulation example. The thresholded TFCE plot display the thresholded TFCE statistic in light blue and for the reference the TFCE support in dark blue. LCE when applied to the TFCE signficant clusters correctly identifies that the smaller cluster is an error. The erroneous TFCE cluster (which constitues a clusterwise error) occurs due to the large area of TFCE support around the image (as shown in dark blue). *[SD: Add a signal outline/transparent image to the cluster images.]*

5.2 Real Data Analyses

In order to illustrate the improvements in detection the LCE can bring we apply to two brain imaging datasets.

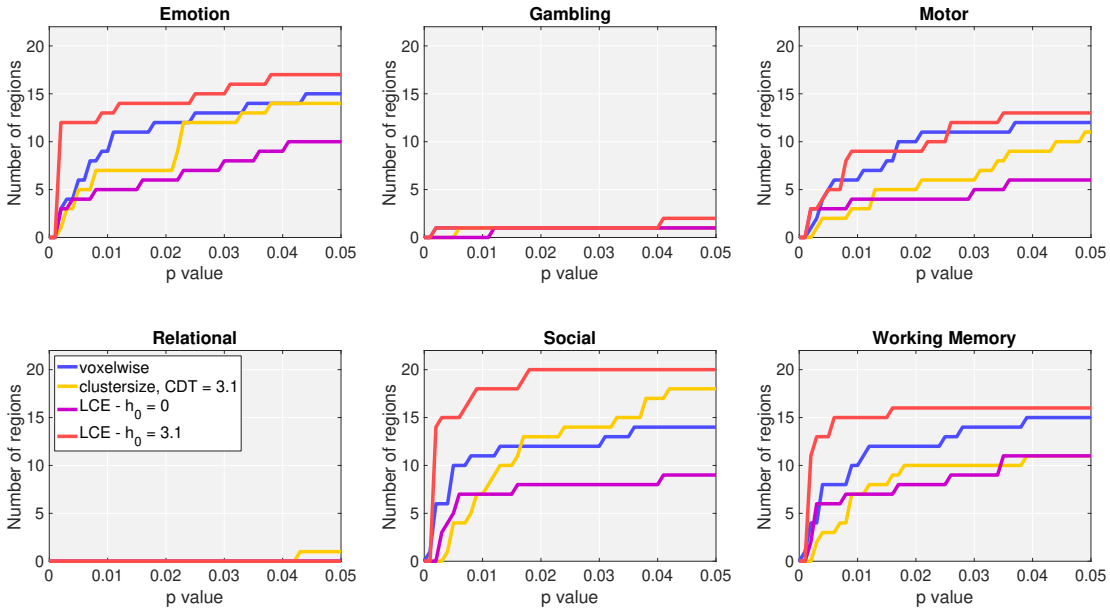

Figure 7: HCP 20 *[SD: vary the line style between the different plots.]*

5.3 Regional Inference

5.3.1 Regional Detection Power

Figures 7 and 8 compares the cumulative number of regions identified as significant at varying p-value thresholds across six cognitive tasks: Emotion, Gambling, Motor, Relational, Social, and Working Memory from the Human Connectome Project dataset. The methods compared include voxelwise inference, cluster size inference with a cluster-defining threshold (CDT) of 3.1, and LCE with two parameter settings ($h_0 = 0$ and $h_0 = 3.1$).

LCE with $h_0 = 3.1$ detects the highest number of regions across all tasks, especially at more lenient p-value thresholds. LCE with $h_0 = 0$ instead shows slightly lower sensitivity. This occurs the threshold $h_0 = 0$ combines information from the signal and the noise which makes it more difficult to distinguish between them.

Voxelwise inference consistently detects more regions compared to cluster size inference, particularly at more liberal p-value thresholds ($p > 0.02$).

The results of LCE applied to the regions of the Harvard Oxford atlas are shown in Figure XXX for the 6 contrasts of the HCP dataset.

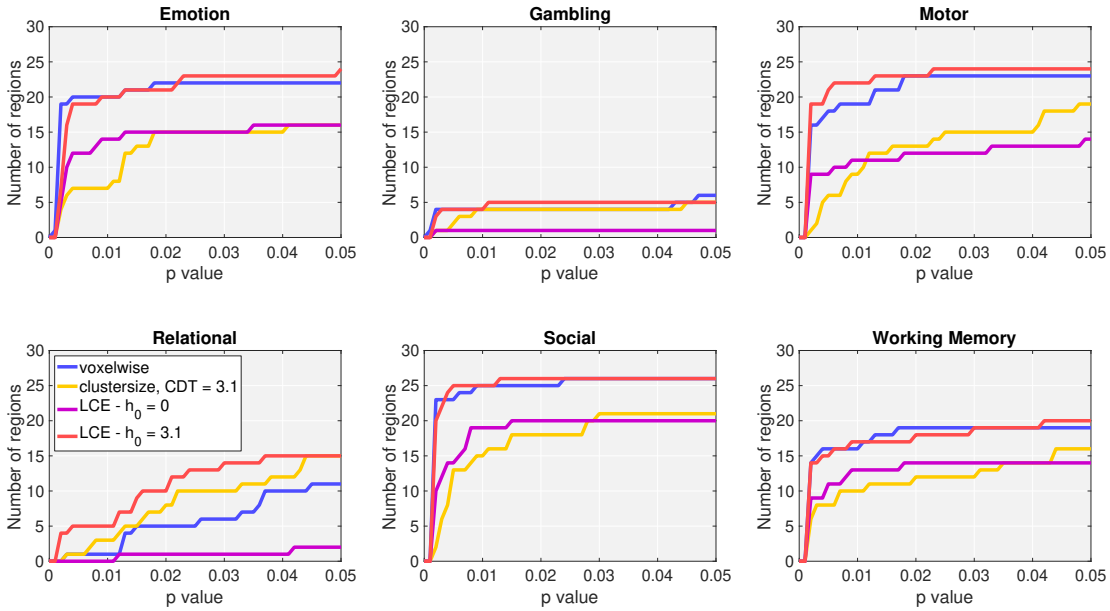

Figure 8: HCP 40

5.3.2 Regional Inference Application

Figure 9 demonstrates the effect of the threshold parameter h_0 on regional inference using LCE applied to anatomically-defined regions from the Harvard-Oxford atlas. The figure displays significant regions detected at $\alpha = 0.05$ for two parameter settings: (a) $h_0 = 0$ (the default TFCE setting) and (b) $h_0 = 3.1$ (aligned with cluster size inference conventions).

The comparison reveals a striking difference in sensitivity between the two approaches. Panel (a) shows that LCE with $h_0 = 0$ identifies only a small number of significant regions, primarily in bilateral superior frontal areas (shown in magenta). In contrast, panel (b) demonstrates that LCE with $h_0 = 3.1$ detects substantially more widespread activation, including not only the superior frontal regions but also extensive bilateral occipital cortex and additional frontal areas (shown in light blue).

5.3.3 Flanker data

[WW: add details] Localizing the inference as described in Section XXX and using the Harvard-Oxford atlas to define the regoins, cluster size inference discovers 14 active while TFCE discover 18. In this example localized TFCE identifies more active regions than cluster size, unlike with the HCP analysis.

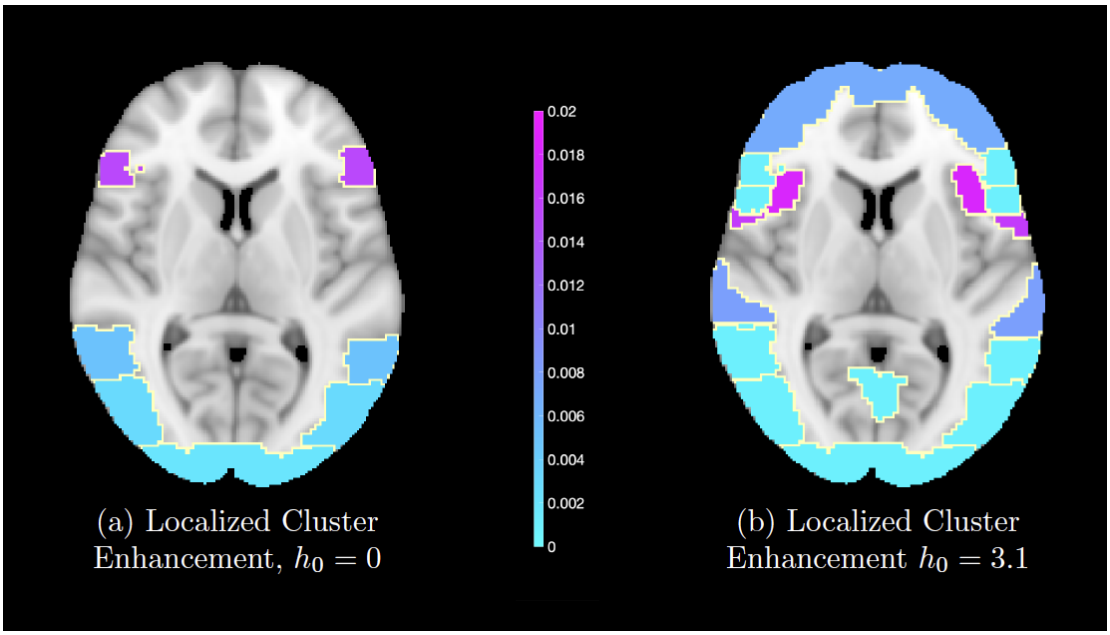

Figure 9: Comparing significant regions for LCE with $h_0 = 0$ and 3.1. These results illustrate that the choice of of $h_0 = 0$ - the default in TFCE - is suboptimal. And we recommend using a higher threshold as used in cluster size inference.

5.4 Cluster-based inference

5.5 Voxelwise inference

6 Discussion

In this paper we have critically evaluated Threshold-Free Cluster Enhancement (TFCE), a widely adopted method in neuroimaging analysis, and introduced Localized Cluster Enhancement (LCE) as a principled improvement. Our work makes three main contributions: (1) we demonstrate that TFCE does not provide valid error control are not supported by theory or simulations, (2) we characterize precisely what error control TFCE does provide, showing it is weaker than cluster size inference, and (3) we propose LCE, which embeds TFCE into a closed testing framework to achieve valid clusterwise and regional error control while maintaining high sensitivity.

6.1 Reassessing TFCE's Claims

6.1.1 Threshold free

As we have seen, despite it's name TFCE is not in fact threshold free. The parameter h_0, which defines the lower bound of integration in equation (1), acts as a de facto cluster-defining threshold. This threshold

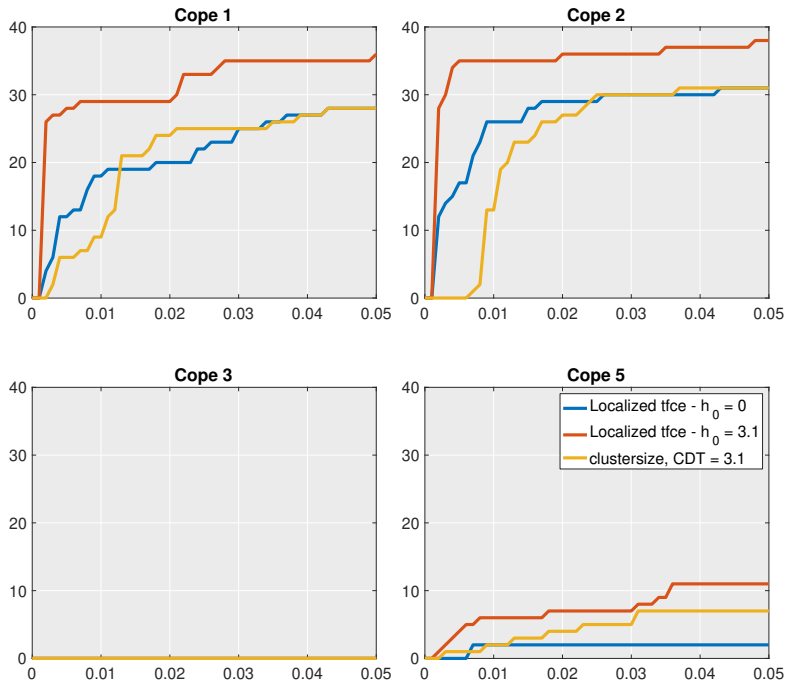

Figure 10: Flanker Copes

curcially determines the spatial support over which TFCE provides valid inference—exactly analogous to how the cluster-defining threshold operates in cluster size inference. In our view this is, in fact, not a drawback of the method, since setting a reasonable threshold can help filter out noise. However, it is important to be explicit about this and to choose the threshold wisely.

Our results in Section 5.2 demonstrate that the default choice is suboptimal. The default choice of $h_0 = 0$ forces the TFCE statistic to integrate across the entire range of test-statistic values, including regions where the signal is indistinguishable from noise. This dilutes the ability to detect genuine activation patterns within anatomically-defined regions. When we applied LCE with $h_0 = 3.1$, we consistently detected more significant regions across multiple tasks in the HCP dataset compared to the default $h_0 = 0$ setting. This suggests that the neuroimaging community should reconsider the default parametrization of TFCE. By setting a positive threshold for the lower integration bound, we exclude low-intensity noise that can obscure true signal.

For researchers who require truly threshold-free inference with strong guarantees, we recommend considering methods like All-Resolutions Inference (ARI) (Rosenblatt et al., 2018), its permutation variant (pARI, Andreella et al., 2023), or the NoTip procedure (Blain et al., 2022). These methods allow valid inference on clusters defined at any threshold, including thresholds chosen after seeing the data.

6.1.2 Spatial specificity and error control

Through simulations (Figure 3) and theoretical analysis (Corollary 2), we demonstrated that TFCE does not control voxelwise error rates as originally claimed in SN. More surprisingly, it also fails to provide clusterwise error control—a weaker form of control that cluster size inference provably achieves.

The root cause of these error control failures is that TFCE violates a key assumption underlying the max-T permutation procedure. The TFCE statistic at an inactive voxel v can depend on data from active voxels through the extent term $e_v(h)$, which measures cluster size across the entire brain. This dependency invalidates the exchangeability assumption required for the single-step max-T method to control voxelwise error rates (Westfall and Young, 1993; Goeman and Solari, 2010).

What TFCE actually controls, as we showed in Corollary 2, is the probability of making any false discovery within the support regions defined at the threshold h_0. As illustrated in Figure 2, these support regions can be extremely large and can indeed span much of the brain. A given significant TFCE cluster is only guaranteed to contain at least one active voxel exists somewhere within its (potentially vast) support region—not within the cluster itself. This represents weaker localization than cluster size inference provides. Indeed one further advantage of raising the threshold h_0 is that the support regions become smaller meaning that discovered activation is more localized.

6.2 Localized Cluster Enhancement as a Solution

LCE addresses TFCE's shortcomings by embedding it within a closed testing framework (Marcus et al., 1976; Goeman and Solari, 2011). This approach, along the lines of other recent improvements to cluster size inference (Goeman et al., 2023), allows us to make valid inferential statements about data-driven regions while maintaining strong error control.

6.2.1 Theoretical advantages

The key innovation of LCE is the use of masked data when testing specific regions. By recalculating TFCE statistics after masking out voxels outside the region of interest (Algorithm 1), we ensure that the test statistic for a null region depends only on data from that null region. This restores the exchangeability properties required for valid permutation inference.

Theorem 4 establishes that LCE provides simultaneous validity over all regions: a much stronger guarantee than TFCE's weak FWER control. Proposition 5 shows how LCE can be used to validate TFCE-identified clusters, correctly identifying which clusters represent true signal versus noise (Figure 6). LCE enables voxelwise inference which is valid simultanouesly together with LCE regional and cluster-based inference - limiting the risk of p-hacking.

6.2.2 Practical advantages

Beyond theoretical validity, LCE offers practical advantages for neuroimaging research. The ability to test regions (either anatomically-defined or user specified) with proper error control enables researchers

to make precise statements about activation patterns. For instance, rather than reporting a large cluster spanning multiple anatomical structures, researchers can identify which specific structures show significant activation.

Our analyses of HCP data (Figures 7 and 8) demonstrate that LCE with $h_0 = 3.1$ consistently detects more significant regions than either voxelwise inference or cluster size inference across six cognitive tasks. This increased sensitivity comes without sacrificing validity—a rare combination in multiple testing procedures. The improved performance stems from LCE's ability to borrow strength across voxels within defined regions while maintaining appropriate error control.

The computational overhead of LCE is minimal. The main TFCE statistics and their permutation distribution need only be calculated once. Testing additional regions requires only recomputing masked TFCE statistics, which is fast. This makes LCE practical for exploratory analyses where researchers may want to test many different regional hypotheses.

6.3 Future Directions

While LCE enables true discovery proportion (TDP) inference in principle (Goeman et al., 2021, since it is a closed testing procedure,), developing practical algorithms for TDP bounds with TFCE-style statistics is not trivial. The extensive spatial borrowing in TFCE makes it difficult to count the minimum number of true discoveries, unlike cluster size inference where efficient algorithms exist (Goeman et al., 2023). Developing such an approach for TDP based inference is an interesting direction for future research.

6.4 Broader Implications

Our work has implications beyond the specific methods discussed. First, it highlights the importance of distinguishing between weak and strong FWER control in neuroimaging. Many proposed methods establish weak control (error rate under the global null) but fail to provide the strong control (error rate under any partial null configuration) needed for proper interpretation. Claims of error control should be accompanied by precise statements about which error rates are controlled and under what assumptions.

Second, our results demonstrate the general value of the use of closed testing frameworks in neuroimaging, building on Goeman et al. (2023). By embedding existing methods like TFCE into closed testing procedures, we can prove and often strengthen their inferential properties without sacrificing sensitivity. This general principle may prove useful for improving other neuroimaging methods.

Third, the discrepancy between TFCE's claims and its actual properties illustrates the need for more rigorous validation of proposed methods. Simulation studies alone are insufficient—formal proofs are necessary to establish error control properties, especially when dependencies between test statistics can violate standard assumptions.

A Proofs of the theorems

In this appendix, we will derive LCE as a closed testing embedding of TFCE. This construction can be used to prove all the theoretical results in the paper.

A.1 Proof of Theorems 2 and 3

Let Π be the group of permutations. Choose any permutation $\tilde{\pi}$. Since $H_{\mathcal{R}}$ implies that $\mathcal{R} \subseteq \mathcal{I}$, by Assumption 1, we have

$$\{S_v\big((\mathbf{X}_{\mathcal{R}} \circ \tilde{\pi}) \circ \pi\big)\}_{v \in \mathbf{R}, \pi \in \Pi} \stackrel{d}{=} \{S_v(\mathbf{X}_{\mathcal{R}} \circ \pi)\}_{v \in \mathbf{R}, \pi \in \Pi},$$

so that the conditions of Theorem 1 of Hemerik and Goeman (2018) are fulfilled. The statement of Theorem 3 now follows immediately from that theorem. Theorem 2 follows by applying Theorem 3 to the case that $\mathcal{R} = \mathcal{B}$.

A.2 A closed testing procedure

We embed the test of Theorem 3 as a local test in a closed testing procedure (Marcus et al., 1976; Genovese and Wasserman, 2006; Goeman and Solari, 2011; Goeman et al., 2021). For every $\mathcal{R} \subseteq \mathcal{B}$, let $\phi_\alpha(\mathcal{R}) = 1$ if $\max_{v \in \mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) > t^\alpha_{\mathcal{R}}$, i.e., if the local test of Theorem 3 rejects the corresponding $H_{\mathcal{R}}$, and $\phi_\alpha(\mathcal{R}) = 0$ otherwise. The resulting closed testing procedure rejects $H_{\mathcal{R}}$ for all $\mathcal{R}$ with $\bar{\phi}(\mathcal{R}) = 1$, where

$$\bar{\phi}_\alpha(\mathcal{R}) = \min_{\mathcal{R} \subseteq \mathcal{S} \subseteq \mathcal{B}} \phi_\alpha(\mathcal{S}).$$

From closed testing theory, we have the following lemma

Lemma 1. $\mathrm{P}(\bar{\phi}_\alpha(\mathcal{R}) = 0 \textit{ for all } \mathcal{R} \subseteq \mathcal{I}) \geq 1 - \alpha$.

Proof. By Theorem 3, $\mathrm{P}(\bar{\phi}_\alpha(\mathcal{R}) = 0 \text{ for all } \mathcal{R} \subseteq \mathcal{I}) \geq \mathrm{P}(\phi_\alpha(\mathcal{I}) = 0) \geq 1 - \alpha$. □

Since $\bar{\phi}_\alpha$ is expensive to calculate, we replace it by a shortcut $\tilde{\phi}_\alpha$, defining $\tilde{\phi}_\alpha(\mathcal{R}) = 1$ if $\max_{v \in \mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) > t^\alpha_{\mathcal{B}}$, and $\tilde{\phi}_\alpha(\mathcal{R}) = 0$ otherwise. We have

Lemma 2. $\tilde{\phi}_\alpha(\mathcal{R}) \leq \bar{\phi}_\alpha(\mathcal{R})$.

Proof. Suppose $\tilde{\phi}_\alpha(\mathcal{R}) = 1$. Choose any $\mathcal{R} \subseteq \mathcal{S} \subseteq \mathcal{B}$. Then

$$\max_{v \in \mathcal{S}} S_v(\mathbf{X}_{\mathcal{S}}) \geq \max_{v \in \mathcal{R}} S_v(\mathbf{X}_{\mathcal{R}}) \geq t^\alpha_{\mathcal{B}} \geq t^\alpha_{\mathcal{S}},$$

so $\phi_\alpha(\mathcal{S}) = 1$. Since $\mathcal{R} \subseteq \mathcal{S} \subseteq \mathcal{B}$ was arbitrary, we have $\bar{\phi}_\alpha(\mathcal{R}) = 1$. □

It follows that

Corollary 3. $\mathrm{P}(\tilde{\phi}_\alpha(\mathcal{R}) = 0 \textit{ for all } \mathcal{R} \subseteq \mathcal{I}) \geq 1 - \alpha$.

A.3 LCE p-values and proof of Theorem 4

The test $\tilde{\phi}_\alpha$ can be equivalently expressed using the LCE p-values defined in (4). We have

Lemma 3. $\tilde{\phi}_\alpha(\mathcal{R}) = 1$ *if and only if* $\mathrm{LCE}(\mathcal{R}) \leq \alpha$.

Proof. We have $\mathrm{LCE}(\mathcal{R}) \leq \alpha$ if and only if there are fewer than αP among

$$\max_{v \in \mathcal{B}} S_v(\mathbf{X} \circ \pi_1), \ldots, \max_{v \in \mathcal{B}} S_v(\mathbf{X} \circ \pi_P) \tag{7}$$

that exceed $\max_{v \in \mathcal{R}} S_v(\mathbf{X}_\mathcal{R})$. To be precise, there are fewer than $P - k$, with $k = \lceil (1-\alpha)P \rceil$ for which this holds. This is in turn equivalent to saying that the kth smallest of (7), i.e., $t^\alpha_\mathcal{B}$, is lower than $\max_{v \in \mathcal{R}} S_v(\mathbf{X}_\mathcal{R})$, which is equivalent to saying that $\tilde{\phi}_\alpha(\mathcal{R}) = 1$. □

Theorem 4 now follows by combining Lemma 3 with Corollary 3.

A.4 Proof of Proposition 5

We now show that every TFCE-"significant" cluster is contained in an LCE-significant cluster and vice versa. Formally, a TFCE-"significant" region is a region of TFCE-"significant" voxels, i.e., voxels v for which $S_v(\mathbf{X}_\mathcal{B} \geq t^\alpha_\mathcal{B}$. A TFCE-"significant" cluster is a TFCE-"significant" region that is connected. An LCE-significant region is a region $\mathcal{R}$ for which $\tilde{\phi}_\alpha(\mathcal{R}) = 1$. A z-thresholded cluster $\mathcal{R}$ is a region for which all voxels v have $T_v \geq z$, and for which all neighbors have $T_v < z$ or are inside $\mathcal{R}$.

Lemma 4. *For every TFCE-"significant" cluster $\mathcal{R}$, there exists an h_0-thresholded cluster $\mathcal{R}' \supseteq \mathcal{R}$ that is LCE-significant. For every LCE-significant region $\mathcal{R}'$, there exists a TFCE-"significant" region $\mathcal{R}$ such that $\mathcal{R} \subseteq \mathcal{R}'$.*

Proof. Let $\mathcal{R}$ be TFCE-"significant". Then all $v \in \mathcal{R}$ have $S_v(\mathbf{X}_\mathcal{B}) \geq t^\alpha_\mathcal{B}$. By the assumptions on f and g, we have $S_v \geq 0$ for all voxels for all permutations, so $t^\alpha_\mathcal{B} > 0$. Since $e_v(h) = 0$ when $T_v < h$, and $g(0) = 0$, $S_v(\mathbf{X}_\mathcal{B}) \geq t^\alpha_\mathcal{B}$ implies that $T_v \geq h_0$. Since $\mathcal{R}$ is connected, there exists an h_0-thresholded region $\mathcal{R}' \supseteq \mathcal{R}$. Choose any $v \in \mathcal{R}'$ and any $h \geq h_0$. For such v and h we have $e_v(h, \mathbf{X}_\mathcal{R}) = e_v(h, \mathbf{X}_\mathcal{B})$, since $\mathcal{C}_v(h) \subseteq \mathcal{R}$, which is because the surroundings of $\mathcal{R}$ have $T_v < h_0 \leq h$. Therefore $S_v(\mathbf{X}_\mathcal{R}) = S_v(\mathbf{X}_\mathcal{B})$. It follows that there exists $v \in \mathcal{R} \subseteq \mathcal{R}'$ for which $S_v(\mathbf{X}_\mathcal{R}) \geq t^\alpha_\mathcal{B}$, so $\mathcal{R}'$ is LCE-significant. This proves the first statement.

To prove the second statement, let $\mathcal{R}'$ be any LCE-significant region. Then there exists $v \in \mathcal{R}'$ such that

$$S_v(\mathbf{X}_\mathcal{B}) \geq S_v(\mathbf{X}_\mathcal{R}) \geq t^\alpha_\mathcal{B},$$

so v is TFCE-"significant", so there exists a TFCE-"significant" region within $\mathcal{R}'$. □

A.5 Voxelwise FWER control using LCE

A voxel v is LCE-voxelwise significant if $\tilde{\phi}(\{v\}) = 1$. For such voxels, the voxelwise FWER property holds, which follows directly from Corollary 3:

Corollary 4. $\mathrm{P}(\tilde{\phi}_\alpha(\{v\}) = 0 \textit{ for all } v \in \mathcal{I}) \geq 1 - \alpha$.

Voxelwise significance is easily characterized by the following Lemma.

Lemma 5. $\tilde{\phi}(\{v\}) = 1$ *if and only if* $g(1)\int_{h_0}^{T_v} f(h) \geq t^\alpha_{\mathcal{B}}$. *For the TFCE setting of* $g(x) = x^{1/2}$, $f(x) = x^2$ *this is equivalent to*

$$T_v \geq (3t^\alpha_{\mathcal{B}} + h_0^3)^{1/3}.$$

Proof. We have that $e_v(h, \mathbf{X}_{\{v\}})$ is simply the indicator that $T_v \geq h$. Therefore, we have $S_v(\mathbf{X}_{\{v\}}) = \int_{h_0}^{T_v} f(h)g(1) + \int_{T_v}^{\infty} f(h)g(0)$, and we note that $g(0) = 0$ by assumption. With TFCE's setting we get

$$g(1)\int_{h_0}^{T_v} f(h) = \int_{h_0}^{T_v} h^2 = \frac{1}{3}(T_v^3 - h_0^3),$$

and the result follows by simple algebra. □

References

Andreella, A. et al. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*.

Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492.

Davenport, S. (2024). StatBrainz matlab toolbox.

Friston, K., Worsley, K. J., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *HBM*, 1:214–220.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417.

Goeman, J. J., Górecki, P., Monajemi, R., Chen, X., Nichols, T. E., and Weeda, W. (2023). Cluster extent inference revisited: quantification and localisation of brain activity. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1128–1153.

Goeman, J. J., Hemerik, J., and Solari, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Annals of Statistics*, 49(2):1218–1238.

Goeman, J. J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics*, pages 3782–3810.

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

Hayasaka, S. and Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *Test*, 27(4):811–825.

Li, H., Nickerson, L. D., Nichols, T. E., and Gao, J.-H. (2017). Comparison of a non-stationary voxelation-corrected cluster-size test with TFCE for group-Level MRI inference. *Human Brain Mapping*, 38(3):1269–1280.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796.

Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., Behrens, T., Johansen-Berg, H., Bannister, P., Luca, M. D., Drobnjak, I., Flitney, D., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J., and Matthews, P. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23 Suppl 1:S208–19.

Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98.

Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., and Kincses, T. (2019). Probabilistic tfce: A generalized combination of cluster size and voxel intensity to increase statistical power. *Neuroimage*, 185:12–26.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419.