Received: 1 September 2020 | Revised: 20 February 2021 | Accepted: 21 February 2021

DOI: 10.1002/hbm.25399

RESEARCH ARTICLE

WILEY

Improving the sensitivity of cluster-based statistics for functional magnetic resonance imaging data

Linda Geerligs | **Eric Maris**

Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Correspondence
Eric Maris, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands.
Email: e.maris@donders.ru.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 451-16-013

Abstract

Because of the high dimensionality of neuroimaging data, identifying a statistical test that is both valid and maximally sensitive is an important challenge. Here, we present a combination of two approaches for functional magnetic resonance imaging (fMRI) data analysis that together result in substantial improvements of the sensitivity of cluster-based statistics. The first approach is to create novel cluster definitions that optimize sensitivity to plausible effect patterns. The second is to adopt a new approach to combine test statistics with different sensitivity profiles, which we call the min(p) method. These innovations are made possible by using the randomization inference framework. In this article, we report on a set of simulations and analyses of real task fMRI data that demonstrate (a) that the proposed methods control the false-alarm rate, (b) that the sensitivity profiles of cluster-based test statistics vary depending on the cluster defining thresholds and cluster definitions, and (c) that the min(p) method for combining these test statistics results in a drastic increase of sensitivity (up to fivefold), compared to existing fMRI analysis methods. This increase in sensitivity is not at the expense of the spatial specificity of the inference.

KEYWORDS
cluster inference, false positives, fMRI, nonparametric, randomization, sensitivity, statistics

1 | INTRODUCTION

Functional magnetic resonance imaging (fMRI) is widely used for clinical and basic neuroscience. The statistical analysis of fMRI data is mostly performed in a parametric framework and using cluster-based statistics (Lindquist & Mejia, 2015). Initially, there was a preference for low cluster-defining thresholds (CDTs), which correspond to large voxel-level p-values ($p > .01$), because of their larger sensitivity for detecting small but widespread effects (Woo, Krishnan, & Wager, 2014). However, in recent years, scientists have argued for increasing CDTs (Eklund, Nichols, & Knutsson, 2016; Woo et al., 2014). Two arguments in favor of high CDTs were put forward, one pertaining to the test's voxel-level false alarm (FA; false positive) rate and the other to its brain-level FA rate (the probability of detecting a significant cluster under the null hypothesis for all voxels): (a) Woo et al. (2014) demonstrated that a low CDT resulted in a merging of nearby effect clusters, indicating an inflated voxel-level FA rate and thus a poor spatial specificity, and (b) Eklund et al. (2016) showed that the brain-level FA rate was not controlled at low CDTs, neither for parametric nor for permutation-based inference. For parametric cluster-based inference, the absence of brain-level FA rate control with low CDTs is not very surprising because the parametric reference distribution for the maximum cluster size is only asymptotically valid; it holds for an increasing CDT under a Gaussian random field (GRF;

Friston, Worsley, Frackowiak, Mazziotta, & Evans, 1994). For permutation-based inference, this is more revealing because it demonstrates a violation of the assumption of independent and symmetric errors (Winkler, Ridgway, Webster, Smith, & Nichols, 2014). These observations have reinforced the use of high CDTs. The downside of this practice is a substantial reduction in sensitivity, especially for detecting widespread but weak effects, and this has recently been demonstrated by Noble, Scheinost, and Constable (2020).

The failure to detect true activations due to lack of statistical power leads to low reproducibility of the results. In the current scientific climate, this may be an even more pressing problem than poor FA rate control (Bansal & Peterson, 2018; Button et al., 2013; Cremers, Wager, & Yarkoni, 2017; Lohmann et al., 2018; Szucs & Ioannidis, 2017). One way to improve statistical power is by increasing the study sample sizes, ideally motivated by a formal power analysis. However, despite an increase in the number of studies with large datasets, the median sample sizes in fMRI studies were still below 30 in 2015 (Poldrack et al., 2017), which is well below the sample size needed to detect large effect sizes (Cohen's $d > 0.8$) with confidence (Geuter, Qi, Welsh, Wager, & Lindquist, 2018). Here, we present an alternative way to improve the statistical power of fMRI studies: we will demonstrate that the sensitivity of statistical tests can be substantially increased (up to fivefold) by combining two approaches: (a) creating test statistics that are affected less by physiologically implausible effect patterns, and (b) adopting a new approach for combining test statistics with different sensitivity profiles.

To achieve this goal, we operate within the randomization inference framework. This framework has a number of important advantages over existing statistical frameworks because it allows (a) to prove FA rate control under a relevant null hypothesis (statistical independence between the biological data and the explanatory variable; see further), (b) the use of an arbitrary test statistic, which allows us to select a test statistic solely on the basis of its sensitivity to the effects of interest, and (c) to combine test statistics with different sensitivity profiles (e.g., different CDTs). All these advantages will be illustrated by the simulations and the analyses on empirical data on which we report in this article.

In the remainder of this introduction section, we will (a) provide a recipe for a group-level randomization test for studies with a within-participants design, (b) prove and discuss the FA rate control of this randomization test, and (c) discuss ways to optimally design a test statistic. In the results section, we will use simulations to demonstrate that the use of different test statistics can substantially increase the sensitivity of fMRI data analysis.

1.1 | A recipe for a group-level randomization test for a within-participants study

The randomization inference framework we propose here, tests the null hypothesis of statistical independence between the biological (i.e., fMRI data) and the explanatory variable (i.e., the experimental conditions). Statistical independence involves that, for the biological data of a randomly sampled participant, it does not matter in which experimental condition it is observed.

The randomization test relies on randomization of the explanatory variable across participants. For a within-participant study, the explanatory variable is the order in which the conditions are presented (denoted as "condition order" in the following). The randomization inference framework requires that there are multiple condition orders that reflect the effect of interest. To clarify the steps that are involved in performing the randomization test, we give an example for one specific study (see Figure 1). This example study involves eight trials and two experimental conditions (A and B), and 20 participants are completing both experimental conditions. Importantly, the first step occurs prior to the data collection.

1. The participants are randomly assigned to one of two condition orders. To optimize sensitivity, it is important to select condition orders that are as different as possible, which is realized by complementary condition orders (e.g., [AABBABAB, BBAABABA]).

2. The fMRI data are collected for every participant.

3. The effect of the experimental conditions (A vs. B) is quantified separately within every participant. When analyzing fMRI data, this is commonly done by running a regression on the voxel-specific MR signals in which the Conditions A and B are represented as separate regressors. Contrast images, which reflect the difference between the beta values of the regressors A and B are typically the basis for the quantification of the effect of interest.

4. A test statistic is computed by combining the effects identified in Step 3 across participants. Typically, this is done by computing a T-statistic across the contrast images. However, the randomization framework allows for any other statistic that reflects the difference between Conditions A and B. In the case of cluster-based statistics, clusters are typically identified by applying a CDT and then counting the number of voxels in this cluster or summing its thresholded voxel-level statistics. Usually, multiple clusters are identified, and the test statistic is then taken as the maximum (for thresholding from below) or the minimum (for thresholding from above) of the cluster-level statistics. The randomization framework allows for many variations on this typical way of calculating cluster-based statistics (see Section 1.3).

5 + 6. The randomization *p*-value is calculated for the observed test statistic. This is done using a reference distribution that is obtained by randomly reassigning the participants to one of the two condition orders, while keeping the observed data (i.e., the single subject contrast images) fixed. Participants are randomly reassigned to one of the two condition orders and the maximum/minimum cluster-based statistic is recalculated. Repeating these steps (random reassignment and recalculation) a large number of times results in the randomization distribution, which is the reference distribution of a randomization test.

7. Each of the observed cluster-level statistics can be compared to the reference distribution to obtain a randomization *p*-value. If one of these *p*-values (the smallest one, which corresponds to the maximum/minimum observed cluster-based statistic) is less than the

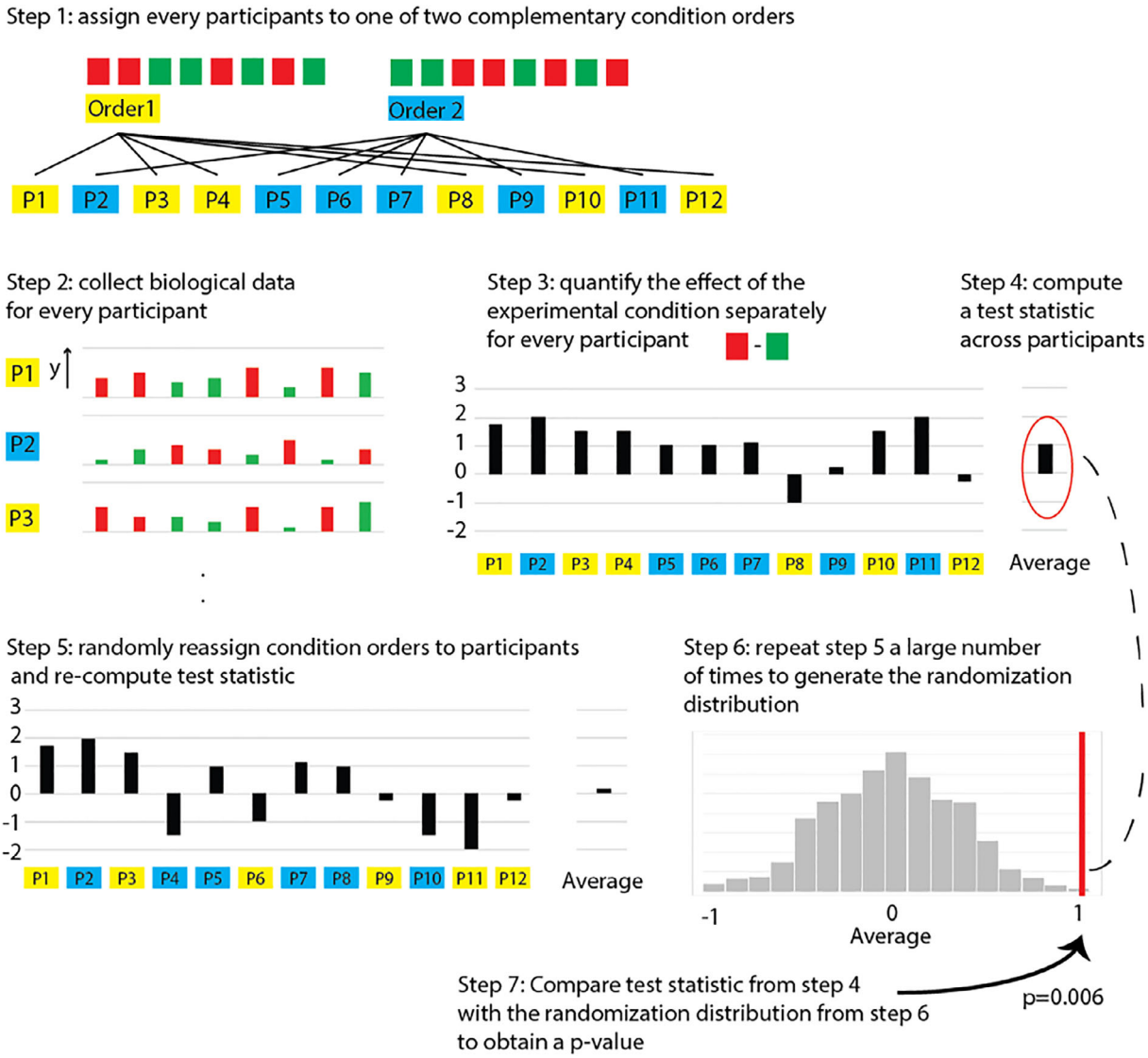


FIGURE 1 Schematic explanation of the steps in the randomization test. This example represents the procedure for a test with one variable of interest (i.e., one voxel). However, the same steps apply in the case of cluster-based statistics, as explained above

nominal alpha level, the null hypothesis of statistical independence between the biological data and the explanatory variable is rejected.

The literature on nonparametric tests for neuroimaging data is dominated by permutation tests (Hayasaka & Nichols, 2003; Hayasaka & Nichols, 2004; Nichols & Holmes, 2002; Pantazis, Nichols, Baillet, & Leahy, 2005; Winkler et al., 2014; Winkler et al., 2016). To describe the difference between a randomization and a permutation test, we must start from a blocked instead of an event-related design. Consider a blocked design with two Conditions A and B, which can be taken in the order AB or BA. Let the data in these two conditions be denoted by the pair $[Y_1, Y_2]$, of which Y_1 is observed in the first condition and Y_2 in the second. Now, with a permutation test, one tests the null hypothesis of exchangeability, which involves that the probability of the data $[Y_1, Y_2]$ is not affected by changing the order of the components (Y_1 and Y_2) over the conditions. In other words, the idea that the identity of the conditions does not matter is captured by the hypothesis that the probability of the pair $[Y_1, Y_2]$ is identical to the pair $[Y_2, Y_1]$. The permutation test compares some statistic of the observed data under the reference distribution that is obtained by randomly permuting the elements of the pair $[Y_1, Y_2]$ separately and independently for all the participants. The essential difference with a randomization test is that, for a permutation test, it does not matter how the participants were assigned to the possible condition orders AB and BA: all to the same condition order, fifty-fifty, random, or fixed. In contrast, for a randomization test, the participants must be randomly assigned to the different condition orders.

1.2 | FA rate control

The randomization test described in the previous subsection controls the brain-level FA rate, and the formal proof is given in the Appendix.

This is not a proof of voxel-level FA rate control, which is a requirement for spatially specific inference. A possible failure to control the voxel-level FA rate is best described for a cluster-based test statistic. Specifically, if one or more of the observed cluster-level statistics exceeds the test statistic's critical value under the randomization distribution, this does not allow for spatially specific statements such as "Voxel X does not belong to a significant cluster, and therefore the probability of an effect at this voxel is less than the nominal FA rate."

Although spatially specific inference is highly useful, we do not consider a formal proof of voxel-level FA rate control a necessary requirement. To our knowledge, there are only two ways to achieve voxel-level FA rate control: (a) Bonferroni correction, and (b) the max(T) test statistic, the maximum of the voxel-level test statistics (Friston, Holmes, Poline, Price, & Frith, 1996). Because of their low sensitivity, both ways are rarely used in practice. A realistic position is to ask for brain-level FA rate control as a first requirement and, only for tests that fulfill this criterium, to evaluate voxel-level FA rate control. We did this in our simulation study: for a number of randomization tests, we not only evaluated their sensitivity, but also their voxel-level FA rate control. It is important to know that not all commonly used statistical tests control the brain-level FA rate. This was shown by Eklund et al. (2016) for cluster-level inference based on GRF theory (Friston et al., 1994) and for cluster-based permutation tests that depend on independent and symmetric errors (Winkler et al., 2014).

We will now describe some differences between our randomization test and parametric statistical tests (e.g., the T- and the F-test) with respect to the nature of the null hypothesis and the auxiliary requirements for a valid statistical test. First, a parametric statistical test controls the FA rate under a null hypothesis that pertains to moments of the probability distribution of the biological data (expected values, variances, regression coefficients, …). Our randomization test, on the other hand, controls the FA rate under the null hypothesis of statistical independence between the biological data and the explanatory variable (the condition orders to which the participants are randomly assigned; in the example above, AABBABAB or BBAABABA). In neuroimaging, researchers typically interpret their effects in terms of the amplitude of the stimulus-evoked hemodynamic response (HR) in relation to the occurrence of particular stimuli. Now, our null hypothesis at the level of the whole biological data are implied by a null hypothesis at the level of the stimulus-evoked HR amplitude: statistical independence between the stimulus-evoked HR amplitude and the experimental Conditions A and B (Maris, 2019). Therefore, by modus tollens, if the latter HR-level null hypothesis is false, then so is the null hypothesis at the level of the whole biological data.

A second difference with a parametric statistical test is that, in its simplest form, our randomization test requires random assignment to one of two condition orders. To maximize sensitivity, we take these condition orders to be each other's complement, but this is not necessary for FA rate control. The requirement of random assignment to only two condition orders can be relaxed by extending the randomization test procedure, and a proof for this extended procedure is given in Maris (2019). For example, consider an existing dataset in which participants were randomly assigned to one of *all possible* condition orders. For this scenario, a valid and sensitive randomization test is also obtained if the randomization distribution is constructed by randomly reassigning every participant to either the (a) condition order to which they were actually assigned, or (b) complement of that particular participant-specific condition order. Thus, every participant has its own pair of complementary condition orders, of which one member is always the observed condition order. Maris (2019) also describes how the randomization test procedure can be extended to allow for designs with more than two conditions, and to explanatory variables that are not under experimental control (e.g., behavioral outcome, non-blood-oxygen-level-dependent [BOLD] physiological variables like EEG and pupil diameter).

A third and last difference with a parametric statistical test is that the latter requires a particular test statistic (e.g., the Z-, T-, or F-statistic). Our randomization test, on the other hand, controls the FA rate for all possible test statistics. In the next subsection, we will make use of this fact to optimize the sensitivity profile of the statistical test.

1.3 | How to construct a test statistic?

We now make use of the fact that the randomization test controls the FA rate for all possible test statistics. This fact allows to construct a test statistic that is maximally sensitive to the effects of interest. There are many ways in which a test statistic for cluster-based inference can be constructed. The first consideration is which voxel-connectivity structure should be used. The connectivity structure determines which voxels should be treated as each other's neighbors, thereby defining the basis for merging voxels in a cluster. Connectivity between voxels can be defined as voxels that share a corner with the current voxel (26 neighbors for each voxel—C26, the FSL default), voxels that share an edge (18 neighbors—C18, the default in SPM) or voxels that share a surface (6 per voxel—C6, the default in AFNI, see Figure 2a). A stricter connectivity definition will result in smaller sized clusters, and this affects the sensitivity for detecting a cluster with some shape of interest. In line with the AFNI defaults, we propose defining neighbors as voxels that are connected via a surface (C6). This will reduce the chance of identifying clusters with voxels that are only connect through a series of corners or edges, as we believe that such a narrow thread is biologically implausible. In our simulation study (see further), we compared the sensitivity of cluster-based test statistics that involve different connectivity definitions.

Another way to vary the cluster definition is by varying the CDT. Strict CDTs are best suited to detect large effects that are present in a small number of voxels. Lenient CDTs are best suited to detect small effects that are present across a large area of the brain. There are also different ways to quantify to the magnitude of a cluster. In the parametric framework, a cluster's magnitude is typically quantified by its size: the number of voxels within the cluster. However, other quantifications such as the sum over the T-values of within-cluster voxels have been shown to be a more sensitive measure in EEG data (Maris & Oostenveld, 2007). Here we use the sum over T-values to determine cluster magnitude.

To devise a potentially sensitive test statistic for fMRI data, we also looked at the cluster patterns in a large number of spatial maps of thresholded T-statistics that were calculated on data without an effect. We observed that especially at low CDTs, there were often two or more separate clusters that were connected via a narrow thread of voxels in between. This resulted in larger cluster sizes in the randomization distribution and less sensitive statistical tests. By adapting the cluster definition, it is possible to avoid such sprawling clusters. One option is to impose a minimum number of above-threshold neighbors that each voxel should have before it is included in the cluster. Here, were we will investigate cluster definitions with no restrictions on the minimum number of neighbors (N0), at least 3 neighbors (N3), 5 neighbors (N5), or 6 neighbors (N6). By removing voxels with a low number of above-threshold neighbors, it is possible to counteract the effects of the data smoothness that is often introduced during preprocessing: voxels at the edge of the cluster will be removed while voxels at the center remain. Single voxels or very small clusters of above threshold voxels are biologically implausible and can be the result of spatial smearing around isolated voxels with high T-values. The effects of these restrictions on the number of neighbors are illustrated in Figure 2b. To further reduce such sprawling clusters, instead of a single removal of voxels with a low number of above threshold neighbors, we can perform this removal several times (denoted as "iterative peeling," with shorthand notation P#, in which the # denotes the number of iterative removals minus one). This way, we can avoid both clusters of isolated voxels, as well as clusters with a small volume that may haphazardly merge to form a larger sprawling cluster. Table 1 provides an overview of the cluster definitions we examined in this article and the shorthand we will use to refer to those definitions in the remainder of the paper.

1.4 | How to combine different test statistics?

A crucial advantage of the randomization framework is that it allows to combine cluster statistics with different sensitivity profiles. A

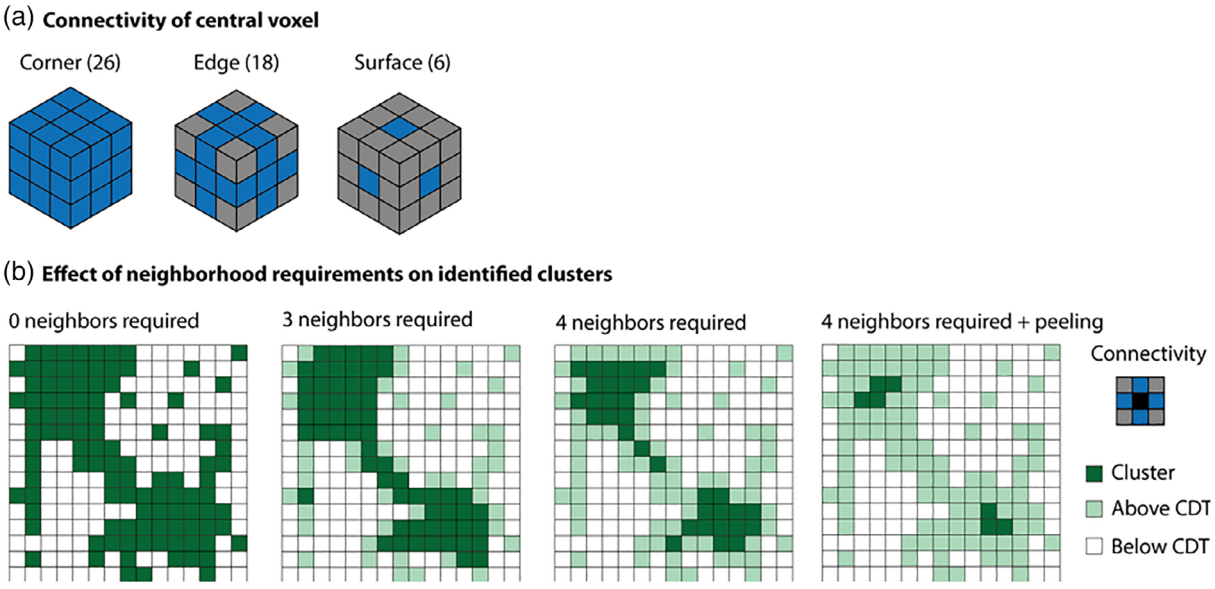


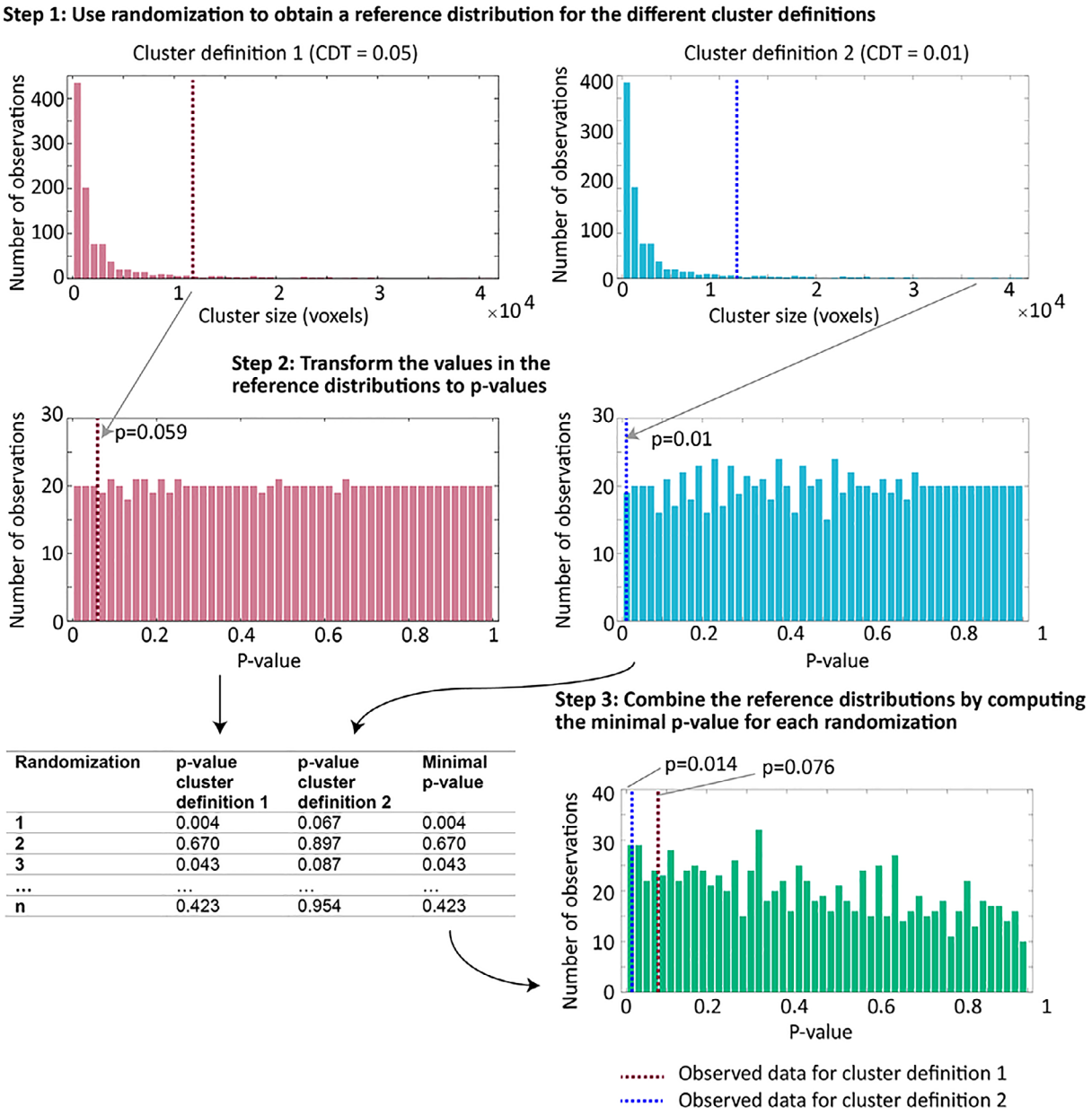
FIGURE 2 (a) Illustration of the three different voxel connectivity structures. Blue voxels are neighbors of the central voxel. (b) 2D illustration of the effect of neighborhood requirements on the identified clusters

TABLE 1 Overview of cluster definitions

Shorthand	Connectivity structure (C)	Neighborhood requirements	Peeling
C18N0P0	18, sharing edges	None	No
C6N0P0	6, sharing surfaces	None	No
C6N3P0	6, sharing surfaces	Min. 3 active neighbors	No
C6N5P0	6, sharing surfaces	Min. 5 active neighbors	No
C6N6P0	6, sharing surfaces	Min. 6 active neighbors	No
C6N6P1	6, sharing surfaces	Min. 6 active neighbors	Apply neighborhood definition in two iterations

Note: Each of these cluster definitions can be paired with different CDTs.
Abbreviation: CDTs, cluster defining thresholds.

researcher may not know whether to expect small or large clusters. In that case, it is possible to analyze the data using different cluster definitions, for example, by varying the CDT, the neighbor definition, and/or the number of required neighboring voxels. Within the randomization framework, the results can be combined over these different cluster definitions. For each of the different cluster definitions (CDT, neighbor definition, etc.), the randomization step results in a distribution of optimum (i.e., maximum or minimum) cluster magnitudes (size or sum). These randomized optimum cluster magnitudes can each be transformed into p-values by comparing them to their corresponding randomization distribution (see Figure 3). By definition, and separately for each of the cluster definitions, the probability distribution of these p-values is uniform (see Figure 3, Step 2). Similarly, each observed cluster magnitude can also be transformed into a p-value by comparing it to its corresponding randomization distribution. After transforming the cluster-definition-specific magnitudes into p-values, these transformed magnitudes can be meaningfully combined in a single randomization distribution. This is realized by taking the minimum p-value over all cluster definitions. This min(p) randomization distribution is constructed in a loop over draws from the randomization distribution: for every draw, evaluate which of the cluster definitions (statistics) has the smallest *p*-value, and use the resulting value as a realization of the min(p) randomization distribution. This min(p) randomization distribution is the final distribution that is used



Randomization	p-value cluster definition 1	p-value cluster definition 2	Minimal p-value
1	0.004	0.067	0.004
2	0.670	0.897	0.670
3	0.043	0.087	0.043
…	…	…	…
n	0.423	0.954	0.423

FIGURE 3 Illustration of how to combine cluster statistics with different sensitivity profiles. This illustration is for combining cluster definitions with different cluster-defining thresholds (CDTs), but the method is the same for combining other test statistics or more than two test statistics

for decision-making: if the observed min(p)-value is less than the $\alpha \times 100$th percentile of the min(p) randomization distribution, then we reject the null hypothesis of statistical independence between biological data and the explanatory variable. By using the min(p) randomization distribution for decision-making (instead of the cluster-definition-specific randomization distributions), we correct for multiple testing (one test per cluster definition).

The min(p) statistic was first proposed by Tippett (1931), but for a different purpose. Pesarin (2001) was the first to propose the min(p) statistic as a component of a nonparametric statistical test, namely as a special case of his nonparametric combination of dependent permutation tests. This method was introduced to neuroimaging by Winkler et al. (2016), but for a different purpose as in the present paper. Winkler et al. (2016) use the min(p) statistic at the level of the single voxels in a situation in which there are multiple statistical tests per voxel. This situation occurs when there are multiple explanatory variables of interest or voxel-level multivariate signals, such as in the case of multimodal imaging and multiple processing pipelines. In this article, we use the min(p) method to combine test statistics that depend on the signal at all voxels jointly, specifically cluster statistics with different sensitivity profiles.

2 | METHODS

2.1 | Data

In our simulation study, we used resting state fMRI data from 103 healthy controls from Oulu dataset in the 1000 Functional Connectomes Project (Biswal et al., 2010) were used for all analyses (http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html). We used the Oulu dataset because previous work showed poor brain-level FA rate control for the permutation tests with this particular dataset (Eklund et al., 2016). The dataset includes 37 male and 66 female participants with a narrow age range (20–23 years, mean = 21.52, *SD* = 0.57). Collection of the data was approved by the ethics committee of the Northern Ostrobothnia Hospital District. Data were collected using a 1.5 Tesla MR scanner, with a repetition time (TR) of 1.8 s. The data consist of 245 time points per subject, $64 \times 64 \times 28$ voxels of size $4 \times 4 \times 4.4$ mm.

For our reanalysis of an existing task dataset, we used fMRI data from 34 participants (25 females; average age of 24.9 years and *SD* of 4.8 years) who participated in an experiment on statistical learning (Richter & de Lange, 2019). The study followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands). Data were collected using a 3 Tesla MR scanner, with a TR of 1 s and a T2*-weighted multiband-6 sequence (TR/TEgeerligs = 1,000/34.0 ms, 66 slices, voxel size 2 mm isotropic, 75° flip angle, A/P phase encoding direction, FOV = 210 mm, BW = 2090 Hz/Px).

In a learning session outside of the MR scanner, participants first learned statistical regularities in object image pairs; in every pair, the second object image was fully predictable based on the first. In the follow-up fMRI session, participants were presented with the same leading object images, but now were followed by the expected trailing object image only in 50% of the cases, and by a different, unexpected trailing object image in the remaining 50%. Participants performed two tasks using these object images: an object categorization task and a character recognition task. In the object categorization task, they categorized the trailing (predictable/unpredictable) object as electronic or nonelectronic, which rendered the object images task-relevant and therefore attended. In the character recognition task, they classified a concurrently shown letter or symbol presented in the fixation dot as a letter or no letter. This task rendered the object images task-irrelevant and therefore unattended. Importantly, the trial order was fully randomized, which made the data suited for applying the randomization test.

2.2 | Simulation design

The aim of the simulations was to investigate the FA rate control and the sensitivity of different test statistics within the randomization framework in comparison to the current standards in the field. We focused on group-level analyses comparing two different task conditions. The resting state data were used in two types of simulations: (a) noise-only simulations using only the resting state data (as performed by Eklund et al., 2016), and (b) simulations in which the resting state data were used as the background signal on top of which we added a stimulus-evoked signal. When the expected magnitude of the stimulus-evoked signal was equal in the two task conditions, we could investigate the brain-level FA rate control of different statistical tests. To investigate the sensitivity profiles of the different statistical tests, we manipulated the (a) between-condition difference in the expected magnitudes of the evoked signals (denoted as "effect size" in the following), and (b) size of the gray-matter volume that exhibited this difference (denoted as "spatial extent" in the following). The effect sizes were quantified as the (population-level) Cohen's *d* of the between-condition differences in the signal magnitudes. The values of Cohen's *d* in our simulation design were 0 (for investigating the brain-level FA rate control), 0.6, 0.8, 1, and 1.2. The task-related BOLD signals were added to the resting state data (see Section 2.3) in a cluster of voxels which was defined by a sphere with a radius of 10, 15, or 20 mm centered at MNI coordinate [3–60 30] (see Figure 4). Thus, our simulation design was 5 (effect size) by 3 (spatial extent).

2.3 | Simulating data

For each of the 15 cells of our simulation design, we performed 1,000 group analyses, each of which started from a random sample of 40 participants from the Oulu dataset. These 1,000 random samples of participants were the same as in Eklund et al. (2016) and were kept constant across all statistical tests.

We simulated an event-related paradigm with two stimulus sequences, of which one was assigned to Condition A and the other

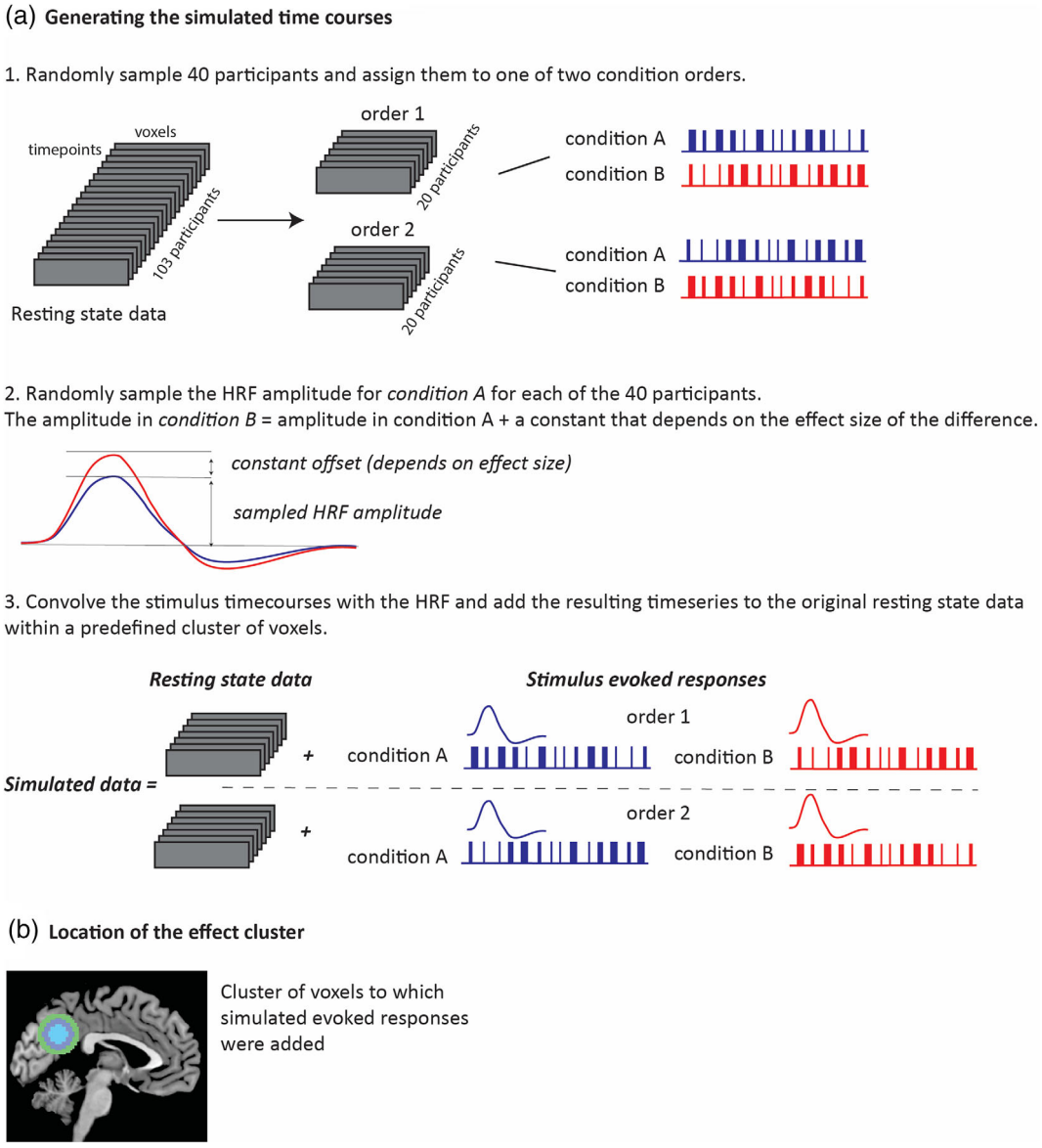


FIGURE 4 (a) Schematic overview of how the data were simulated. (b) Illustration of the sizes of the (unsmoothed) simulated clusters with radii of 10, 15, and 20 mm

to Condition B. Each stimulus sequence had 62 simulated stimulus onsets with random durations (1–4 s) and inter-stimulus intervals (3–6 s; same as E2 in Eklund et al., 2016). The order of stimulus durations and inter-stimulus intervals was reversed for sequence 2 as compared to sequence 1. The same stimulus sequences were used for all participants. In line with the argument in *A recipe for performing a group-level randomization test for a within-participants study*, we simulated data for two complementary condition orders. For one condition order, one sequence was associated with condition A and the other to Condition B. For the complementary condition order, this was reversed. In each random sample of 40 participants, half of the participants were randomly assigned to one condition order and the other half to the complementary order.

The amplitude of the simulated evoked blood-oxygen-level-dependent (BOLD) response varied across participants but was the same for different events within the same individual. For each participant in the Oulu dataset, the amplitude of the evoked response in each of the two task conditions was drawn from a normal distribution with a mean of 4.5 and *SD* of 2.25. Using these values, when contrasting each task condition with the baseline, a significant cluster was found in approximately 90% of 1,000 stimulations (using GRF-based inference with a CDT of $p < .001$). Differences in the evoked responses between task conditions were introduced by adding a constant value to the evoked response amplitudes of all participants for condition A. The size of this constant was chosen such that the effect size of the between-condition difference for the evoked response amplitudes (quantified as Cohen's *d*) was either zero (for testing the FA rate control), 0.6, 0.8, 1, or 1.2.

Two task-related BOLD signals were created for each participant by convolving the stimulus functions (specifying onset and duration)

with the canonical hemodynamic response function (HRF) and multiplying the resulting time course with the amplitude of the evoked response in each task condition. After convolution with the canonical HRF, the regressors for Conditions A and B were uncorrelated ($r < .01$).

The task-related BOLD signals were added to the resting state data in a cluster of voxels which was defined by a sphere with a radius of 10, 15, or 20 mm centered at MNI coordinate [3–60 30] (see Figure 1). To ensure that the assumptions of GRF theory were met, the image containing the cluster definition (specified as zeros and ones) was smoothed with a Gaussian kernel (6 mm FWHM, same as the smoothness of the resting state data, see Section 2.4). When combining the task-related BOLD signal and the resting state data, the task signal was multiplied by the weights in the cluster definition image. True positive voxels were defined as voxels that were part of the cluster definition before smoothing. True negative voxels were defined as voxels that contained less than 1% of the task-related signal (i.e., after smoothing). Only these true positive and true negative voxels were considered in metrics of sensitivity or specificity of the statistical tests.

2.4 | fMRI data analyses

The fMRI data used for the simulations were preprocessed using standard SPM 8 processing pipelines (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), including realignment, coregistration, normalization, and 6 mm FWHM smoothing (see Eklund et al., 2016 for more details). The fMRI data were not corrected for geometric distortions, as no field maps are available.

A general linear model (GLM) was applied to the preprocessed fMRI data, using two regressors for each of the two task conditions (A and B): the HRF-convolved stimulus function (specifying onsets and durations) and its first derivative. The stimulus onset and duration times in the GLM were matched to the stimulus onset and duration times that were used to simulate the data (which depend on the condition order). The estimated head motion parameters were used as additional regressors in the design matrix, to reduce effects of head motion. To account for low-frequency drifts in the data, a discrete cosine transform with cutoff of 128 s was used. Temporal correlations were corrected for with a global AR(1) model in SPM. The first-level contrast between task Conditions A and B (A-B) was used as the input for the group-level analyses.

For the analysis of the task fMRI dataset, we used the first-level contrast images provided by the authors of the original study (Richter & de Lange, 2019). Preprocessing and first-level modeling of the data was done using FSL 5.0.11 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl), as described in Richter and de Lange (2019). The original analyses of these data demonstrated that neural activity was attenuated for expected compared to unexpected stimuli when stimuli were attended, but not when they were unattended (Richter & de Lange, 2019). Here we statistically tested the simple effect of expectation (expected vs. unexpected trailing objects) in the attended condition.

In the group-analyses, we looked for voxels or clusters in which the first level contrast of interest was significantly different from zero. As a baseline for the evaluation of the performance of the randomization test, we investigated the FA rate and the sensitivity of the following alternative statistical methods: thresholding using false discovery rate (FDR) control (Genovese, Lazar, & Nichols, 2002), cluster-level inference using GRF theory (Friston et al., 1994), and control using the randomization distribution of threshold-free cluster enhancement (TFCE; Smith & Nichols, 2009). For the analyses relying on the randomization distribution, we used the maximum value the test statistic of interest (summed within-cluster T-statistics or voxel-specific TFCE values) to correct for multiple comparisons. To display whole-brain results, we used the MATLAB data visualization toolbox Slice Display (Zandbelt, 2017).

2.5 | Replicating and extending our simulation study

For the purpose of replicating and extending our simulation study, we have shared the preprocessed fMRI data and the code (a MATLAB script and a library of functions) that were used to produce the results of this simulation study (see https://doi.org/10.34973/zw83-tn77). The script was shared as a Life Script, a MATLAB format that is specifically designed for documenting code. The library of MATLAB functions is documented by means of extensive help text. Crucially, this function library can also be used for the second-level analysis of one's own fMRI data. The first-level analyses can be performed using any of the existing fMRI analysis packages. If SPM 8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) is used for the first-level analyses, then the shared code requires only minor changes to produce the required input to the second-level functions.

The preprocessed resting state fMRI data and the library of MATLAB functions will be publicly available at https://doi.org/10.34973/zw83-tn77 only after the paper is published. However, for reviewing purposes, the editor can request an anonymous link to this collection, which can then be forwarded to the reviewers. The Life Script is also shared as Supporting Information to this article.

3 | RESULTS

3.1 | An empirical check of the FA rate control

As a part of our simulation study, we performed an empirical check of the correctness of the proof in the Appendix. We ran two types of simulations under the null hypothesis (statistical independence between the biological and the explanatory variable): (a) noise only simulations using raw resting state data (as performed by Eklund et al., 2016), and (b) simulations using simulated fMRI data in which every participant's data exhibited nonzero stimulus-evoked HR amplitudes to both experimental conditions within a restricted cluster of voxels (see Section 2). When used for checking the FA rate control, the expected values of the stimulus-evoked HR amplitudes (calculated over the population of

participants) were equal in the two conditions (see Section 2). We calculated the brain-level FA rates of the cluster definitions in Table 1, each combined with four CDTs (0.05, 0.01, 0.005, and 0.001), and compared these with the brain-level FA rates of three alternative popular methods: FDR control (Genovese et al., 2002), cluster-level GRF inference (Friston et al., 1994), and control using the randomization distribution of TFCE (Smith & Nichols, 2009). FDR and GRF control have a rationale in the parametric framework, and TFCE achieves brain-level FA rate control by making use of the randomization distribution of the maximum TFCE-value (instead of the maximum cluster statistic, as in our approach). TFCE can be performed for different connectivity structures, and here we used the surface (TFCE-C6) and the corner connectivity structure (TFCE-C26). TFCE depends on two tuning parameters, and for the first set of results, we used the values that were also used in the original publication of the method (Smith & Nichols, 2009). In a later set of results, we also report on the FA rate control of randomization tests using min(p) statistics.

The two types of simulations (noise only and noise plus a stimulus-evoked signal) resulted in almost identical results. Figure 5 shows the results for the second type of simulations (stimulus-evoked signals with equal expected values in the two conditions). These results support the proof in the Appendix: the randomization test controlled the brain-level FA rate for each of the cluster definitions in Table 1 and each of the CDTs. FDR and the randomization-based TFCE also controlled the FA rate, although FDR was too conservative. Importantly, for the parametric inference based on GRF theory the brain-level FA rate was only accurately controlled for a CDT equal to 0.001; it rose up to 32% for parametric inference with a CDT of 0.05. Therefore, in the remainder of the paper, we only consider GRF-based inference using a CDT equal to 0.001.

3.2 | Sensitivity

The main objective of our simulation study was to investigate the sensitivity of the different test statistics to detect simulated effects in the data. To this end, we simulated data using stimulus-evoked HR amplitudes of which the expected values differed between the two conditions (see Section 2). In the simulation design, we varied across four effect sizes (Cohen's *d*s of 0.6, 0.8, 1, and 1.2; see Section 2) and three simulated true cluster sizes (spheres with radii of 10, 15, and 20 mm; see Section 2). As a baseline for our comparisons, we calculated the sensitivity of the three alternative popular methods: FDR, GRF, and randomization-based TFCE.

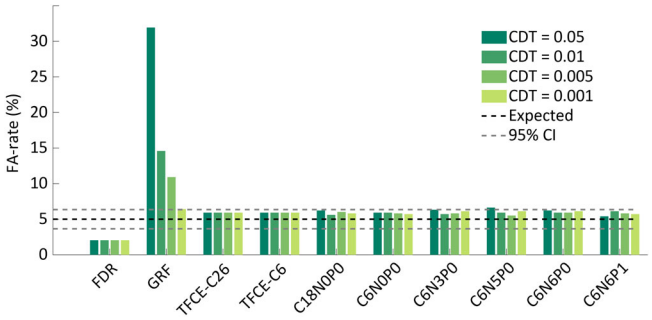


FIGURE 5 The observed brain-level false alarm (FA) rate for the Gaussian random field (GRF) theory and for each of the six different cluster definitions that we used in the randomization testing framework. The dotted lines show the binomial 95% confidence interval around the 5% nominal FA rate for 1,000 simulations

There are several ways of quantifying sensitivity. Here, we quantify sensitivity using a measure that reflects the aggregated voxel-specific hit rates over the brain. Specifically, we calculated the identification rate, which is the proportion of simulations in which the significant clusters (those with a *p*-value less than .05) cover at least half of the simulated cluster. This reflects our interest in identifying the location of the effect, instead of only detecting the presence of an effect somewhere in the brain. At the end of the results section, we will also report on the brain-level hit rate, which does not depend on the coverage of the simulated cluster, but only on whether or not a cluster was significant.

Figure 6 shows that, when the cluster is large or the effect size is small, TFCE shows the highest identification rate. In the other cases, GRF with CDT = 0.001 is the most sensitive test. FDR is the least sensitive test statistic. TFCE with the surface connectivity structure (TFCE-C6) is slightly more sensitive than TFCE with the corner connectivity structure (TFCE-C26). Therefore, we will use TFCE-C6 and GRF with CDT = 0.001 as the reference statistics in the comparison with our cluster statistics.

Next, we compared the sensitivity of the different cluster definitions in Table 1 across the four CDTs. Figure 7 shows that the choice for a particular CDT and cluster-definition has a large impact on the sensitivity of the statistical test. We observed that the sensitivity of the test statistics involves a trade-off between CDT and cluster definition: cluster definitions with more neighborhood restrictions tend to be more sensitive when they are combined with more lenient CDTs, while cluster definitions with fewer restrictions tend to be more sensitive when combined with stricter CDTs. Also, for small effect sizes, the lenient CDTs are always more sensitive than the strict CDTs. On the other hand, for intermediate and large effect sizes, the strict CDTs tend to be more sensitive in the case of a small cluster size while the more lenient CDTs tend to be more sensitive for a large cluster size.

To characterize the sensitivity of the different cluster definitions for the real task fMRI dataset we analyzed, we visualized the randomization distribution of the maximum cluster statistic (Figure 8a). We observed that the distance between the observed maximum cluster statistic and its randomization distribution increased for smaller CDTs (i.e., for more selective criteria) and for stricter neighborhood definitions (especially for C6N6P1). This pattern of results was very similar to our simulations with an intermediate cluster extent and a large effect size.

In Figure 8b, we show how the number of active voxels decreases as a result of decreasing the CDT and increasing the restrictiveness of the neighborhood definition. The effects of the CDT and the restrictiveness of the neighborhood definition on the spatial distribution of the active voxels are illustrated in Figure 8c,d. While for many

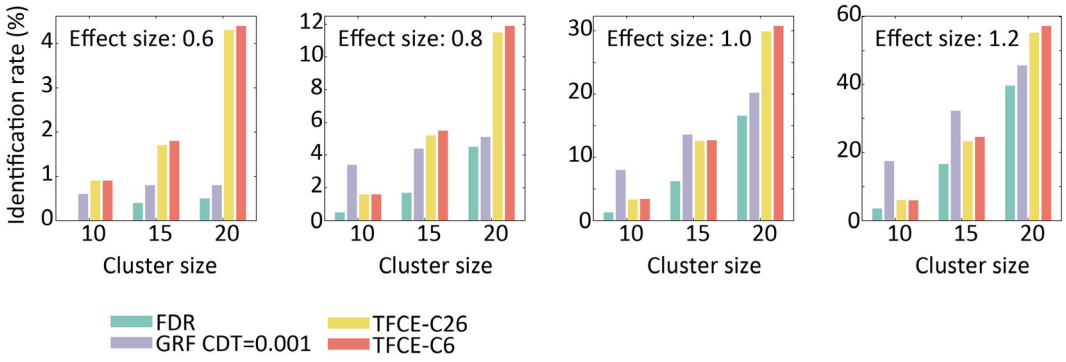


FIGURE 6 The observed identification rates for four test statistics that are commonly used in functional magnetic resonance imaging (fMRI) research. The identification rate is the % of simulations in which a significant cluster overlaps with at least half of the simulated cluster. Note the different scaling of the axes for the four simulated effect sizes, reflecting the fact that the identification rate depends on the simulated effect size

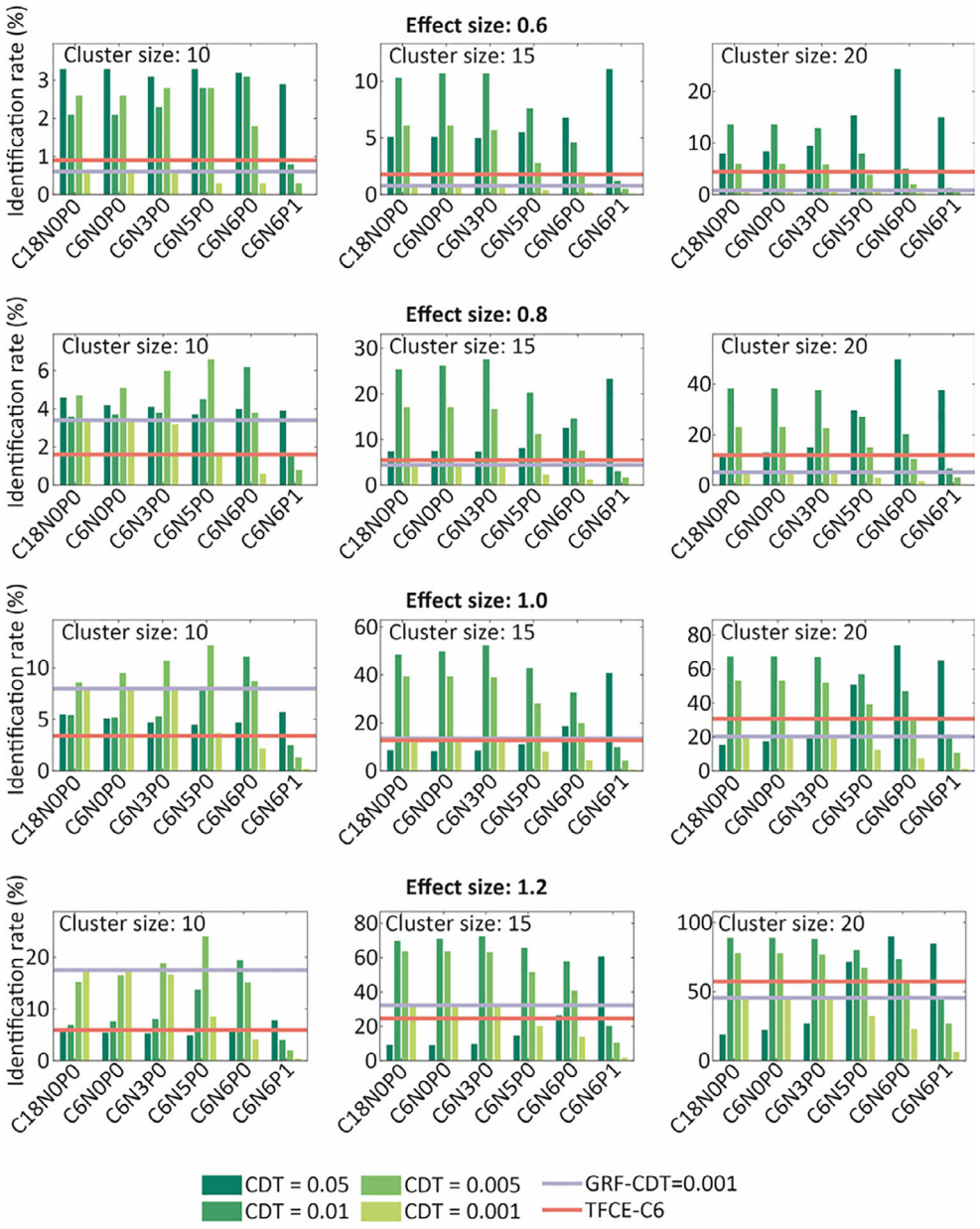


FIGURE 7 The observed identification rates for each of the basic different cluster definitions that we used in the randomization testing framework. As a reference, we also plotted the identification rates for Gaussian random field (GRF) with a cluster-defining threshold (CDT) of $p < .001$ and for TFCE-C6

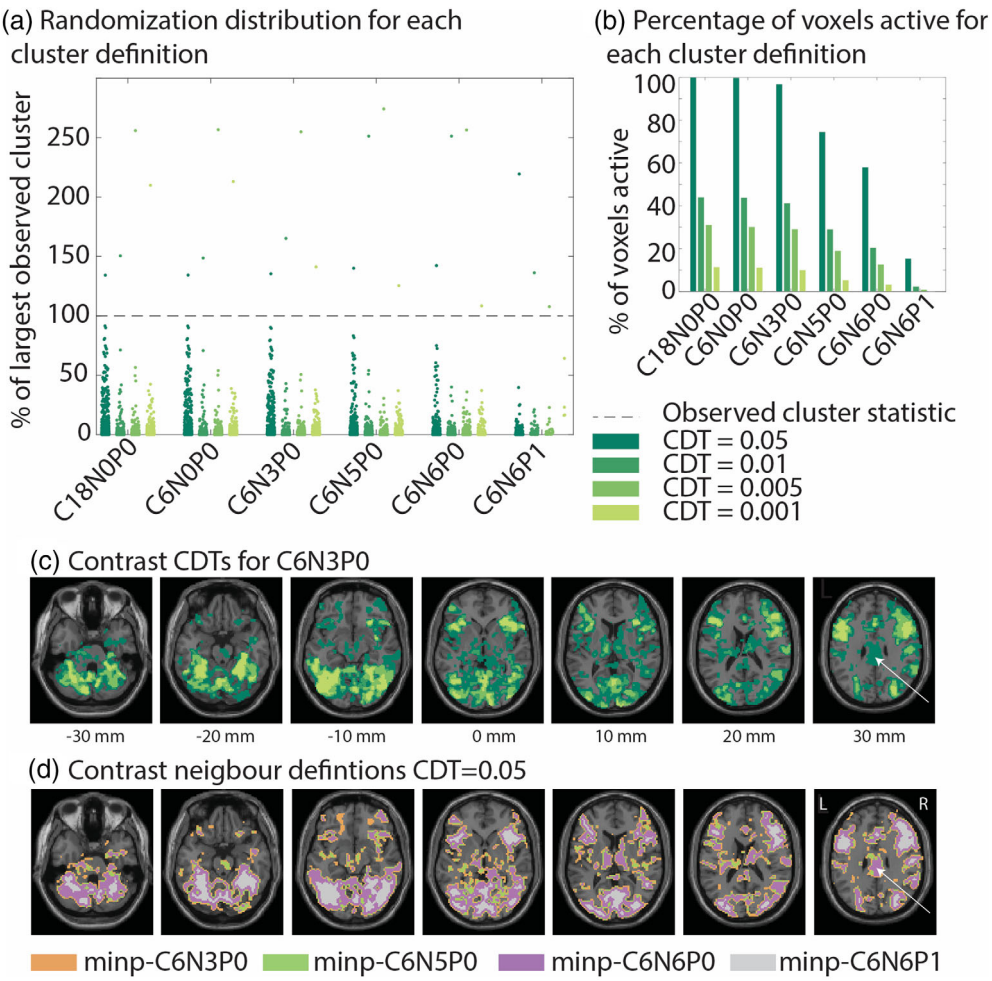


FIGURE 8 Analyses of real task functional magnetic resonance imaging (fMRI) data. (a) Each dot indicates a realization of the randomization distribution of the maximum cluster statistic for a specific cluster definition, scaled as a percentage relative to the maximum observed cluster statistic for that cluster definition. (b) The percentage of active voxels identified in the group analyses, based on different cluster definitions. (c) The voxels that are included in the active clusters depending on the chosen cluster-defining threshold (CDT) (based on C6N3P0). (d) The voxels that are included in the active clusters depending on the chosen neighborhood definition (based on CDT = 0.05). The arrow indicates a region in anterior cingulate where a cluster is preserved with a strict neighborhood definition but not with a strict CDT

clusters, we observe similar effects of decreasing the CDT and increasing the restrictiveness of the neighborhood definition, there are also some interesting differences. For example, in the anterior cingulate gyrus (marked by a white arrow in Figure 8b) the cluster of activation disappears when the CDT decreases but it is preserved when a more restrictive neighborhood definition (C6N6P0) is combined with a larger CDT (CDT = 0.05), due to its large spatial extent.

The three cluster definitions with the fewest constraints on neighborhood structure (C18N0P0, C6N0P0, and C6N3P0) showed very similar sensitivity and activity patterns, and therefore we will only consider C6N3P0 in the remainder of the paper. Crucially, our simulation results show that for all effect sizes and cluster sizes, we can identify at least one test statistic that outperforms both GRF at $p < .001$ and TFCE-C6. However, the sensitivity to different effect sizes and cluster sizes differs widely across test statistics (i.e., cluster-definition-CDT combinations) and it is not possible to choose a single test statistic that performs best in all cases.

3.3 | The effect of combining cluster definitions

Figures 7 and 8 show that different test statistics are optimal for different combinations of effect sizes and cluster sizes. This was our motivation for combining different test statistics by means of the min(p) method (see Section 1.4). In particular, for each of the four remaining cluster definitions (C6N3P0, C6N5P0, C6N6P0, and C6N6P1) we computed a combined test statistic that combines across different CDTs (0.05, 0.01, 0.005, and 0.001). We also used the min(p) method to compute a combined cluster statistic that combines across all of these four cluster definitions and CDTs. The combined test statistics for the four cluster definitions (across all CDTs) are denoted as minp-C6N3P0, minp-C6N5P0, minp-C6N6P0, and minp-C6N6P1, and the combined test statistic across all cluster definitions and CDTs is denoted as minp-all.

Figure 9a shows the identification rates for the minp-C6N3P0 test statistic. Importantly, these identification rates were calculated on the basis of corrected *p*-values for the clusters, that is, the *p*-values were calculated under the randomization distribution of the minp-C6N3P0 test statistic. Figure 9a shows that combining the C6N3P0 test statistic over CDTs using the min(p) method results in a sensitivity that is similar to the best performing CDT, for all cluster sizes and effect sizes. For large and small cluster sizes, the minp-C6N3P0 statistic performs a little better than the best performing CDT, whereas for the intermediate cluster size it performs a little bit worse. Our finding that the minp-C6N3P0 statistic performs as well or nearly as well as the optimal CDT for the C6N3P0 statistic shows that the correction

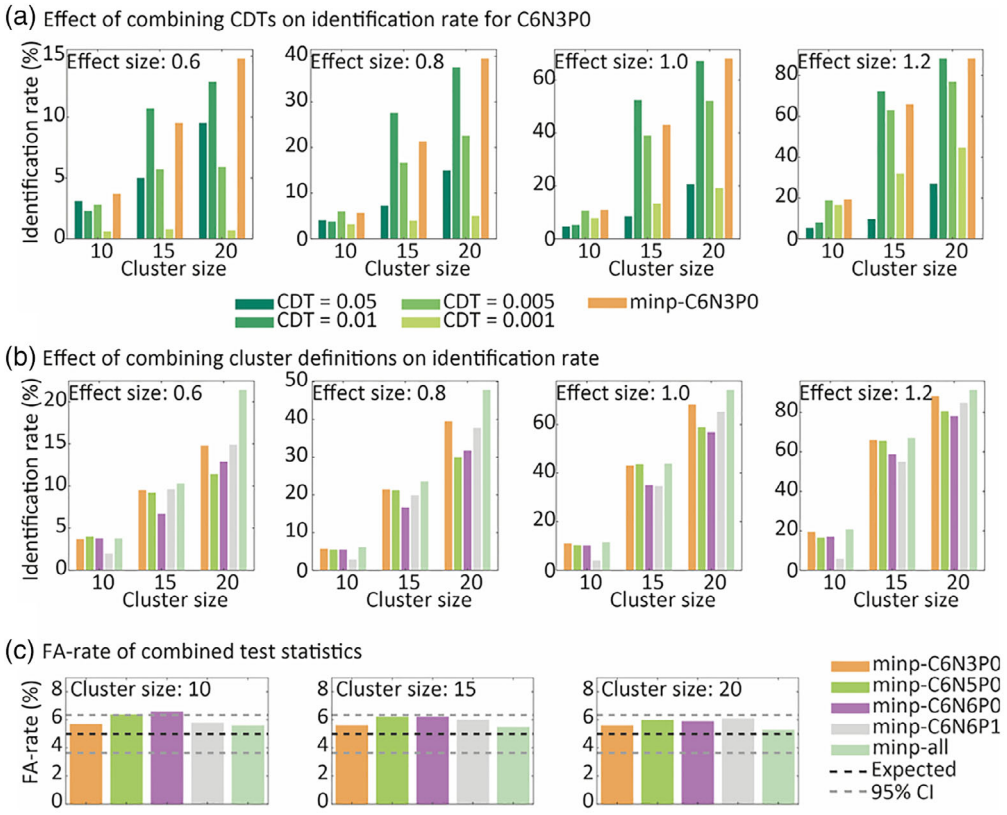


FIGURE 9 (a) The effect of combining cluster-defining threshold's (CDT's) using the min(p) method on the identification rate is illustrated for the C6N3P0 test statistic. (b) The effects of combining cluster definitions using the min(p) method on the identification rate. Each of the cluster definitions was also combined across CDTs using the min(p) method. (c) False alarm (FA) rates of the combined cluster definitions

for multiple testing (i.e., the multiple CDT-specific C6N3P0 statistics) using the min(p) method has only minimal effects on the sensitivity.

The results for the other three cluster definitions (C6N5P0, C6N6P0, and C6N6P1) are highly similar. However, there are substantial differences between the different cluster definitions, as is also clear from Figures 7 and 8. This fact motivates the use of the minp-all statistic, for which the results are shown in Figure 9b. This figure shows that combining different cluster definitions using the min(p) method results in further improvements in sensitivity. In fact, for all cluster sizes and effect sizes, the minp-all statistic is equally sensitive or more sensitive than the best performing single cluster definition statistic. This shows that combining different cluster definitions using the min(p) method results in a better sensitivity for the different types of effect.

As a further check on the proof in the Appendix, we also calculated the empirical brain-level FA rates for the different min(p) test statistics. Figure 9c shows that all these test statistics control the FA rate at their nominal values.

3.4 | Combining TFCE-parameters

In Figure 9, we reported on the results that were obtained by combining different CDTs and different cluster definitions into a single test statistic by means of the min(p) method. The same method can also be used to combine across the different tuning parameter values for TFCE. The calculation of the TFCE image depends on a width and a height parameter, and different values for these parameters may result in a different sensitivity profile of the associated statistical test. Therefore, we also investigated the sensitivity of a min(p)-TFCE statistic that combined across all 25 tuning parameter combinations that were considered in the original paper by Smith and Nichols (2009). We applied this method to TFCE with surface (TFCE-C6) connectivity structure. Figure 10a shows that the min(p)-TFCE-C6 statistic shows better sensitivity than TFCE-C6 for small effect sizes. However, for intermediate and large effects sizes we did not observe an advantage of combining across TFCE parameter settings.

3.5 | Comparing the best test statistics

We now compare the four test statistics that are the most promising on the basis of our previous analyses: GRF with CDT = 0.001, TFCE-C6, minp-TFCE, and minp-all. Figure 10a shows the identification rates for these four test statistics. We found that, for all simulated cluster sizes and effect sizes, minp-all outperformed all three other test statistics (the existing methods). For small effect sizes (0.6 and 0.8) and intermediate or large clusters, the identification rate of minp-all is up to five times larger than the one for the best performing existing method. For larger effect sizes (1.0 and 1.2), the identification rate increases for all test statistics, but minp-all continues to outperform the existing methods. For example, for large clusters with an effect size of 1, the identification rate increases from 31 to 74%.

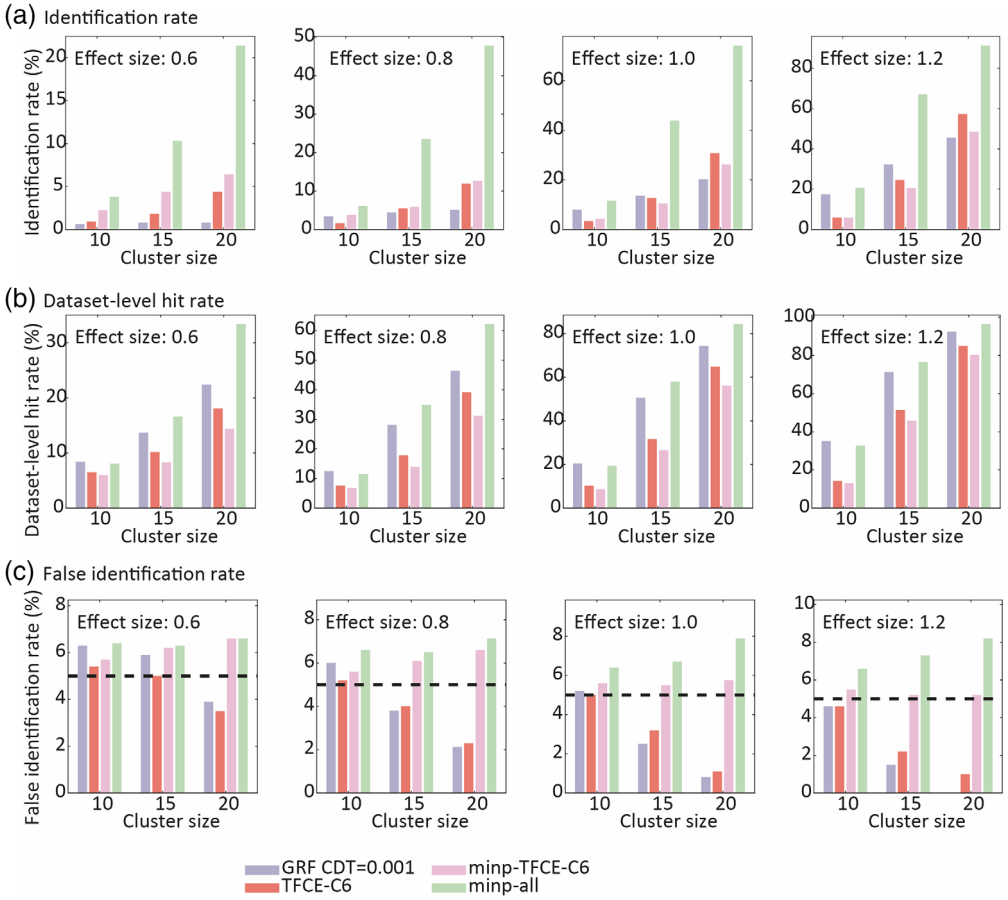


FIGURE 10 Different measures of sensitivity and spatial specificity of the Gaussian random field (GRF), TFCE, minp-TFCE, and minp-all test statistics. (a) The identification rate, which is the % of simulations in which a significant cluster overlaps with at least half of the simulated cluster. (b) The dataset hit rate, which shows the percentage of simulations in which an above threshold cluster was detected, regardless of whether this overlapped with the simulated cluster. (c) The spatial specificity, which is measured as the percentage of simulations in which more false positive than true positive voxels were detected

Identification rate, the measure of sensitivity on which we reported so far, is the probability that at least 50% of the voxels in the simulated cluster is detected as significant. This measure depends on two factors: (a) the probability that one or more clusters are significant (the brain-level hit rate), and (b) the probability that these significant clusters cover more than 50% of the voxels in the simulated cluster. In Figure 10b, we show the brain-level hit rates for the four best test statistics. In terms of this measure, for medium and large clusters, the minp-all statistic outperformed all other test statistics. For small clusters, GRF at CDT = 0.001 slightly outperformed the minp-all statistic. Together, the results in Figure 10a, b suggest that GRF at CDT = 0.001 is good at detecting whether there is an effect, but performs poorly in identifying the spatial extent of the effect. The minp-all statistic performs well for both measures of sensitivity.

The flipside of a better effect coverage is a potentially reduced spatial specificity. To investigate whether this is indeed the case, we calculated the percentage of simulations in which the number of false positive voxels was larger than the number of true positive voxels, which we will call the false identification rate. The result of this analysis is shown in Figure 10c. For small cluster sizes, the false identification rate for all four test statistics is approximately the same. For intermediate and large clusters, the false identification rate is higher for the minp-all statistic than for the existing methods. However, in absolute terms, the minp-all statistic showed adequate spatial sensitivity for all cluster and effect sizes, as the false identification rate was always between 6 and 8%.

In Figure 11, we show the results of our analyses of the real task fMRI data. These analyses pertain to the contrast "unexpected minus expected" and they reflect the suppression of the stimulus-evoked response when the objects expected. Although not relevant for the present paper, this suppression was only observed when the objects were attended. For each of the four statistical tests, we obtained a significant difference between the expected and the unexpected condition. However, there is a substantial variability in the activation maps. In line with our simulation results, we find that the spatial extent for GRF at CDT = 0.001 is very small. The spatial extent for minp-all and TFCE-C6 are very similar and both are much larger than GRF at CDT = 0.001. The spatial extent for minp-TFCE-C6 is the largest and some of the included voxels have a very low t-statistic, with the minimum t-statistic equaling 1.000018. This liberal behavior of the minp-TFCE-C6 statistic seems inconsistent with the fact that, in our simulation study, the minp-all statistic was more sensitive for all combinations of effect size and true effect cluster size. It is possible that the real task fMRI data have true active voxels that can only be detected when the TFCE is performed using appropriate tuning parameters. Alternatively, the larger spatial extent for the minp-TFCE-C6 statistic may be the result of spatial smearing that increases the false identification rate (Woo et al., 2014).

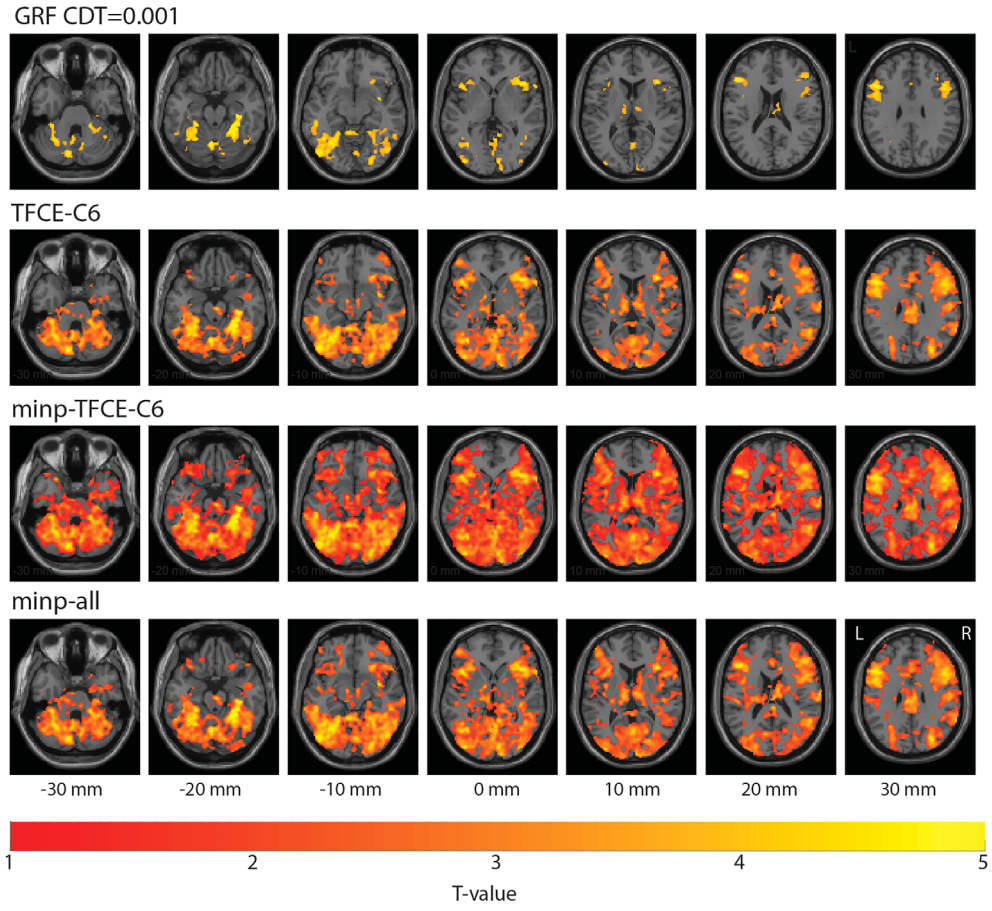


FIGURE 11 Analyses of real task functional magnetic resonance imaging (fMRI) data. Map of significantly active voxels for the Gaussian random field (GRF), TFCE, minp-TFCE, and minp-all test statistics. The colors in each activity map indicate the T-statistic for the contrast "unexpected minus expected"

4 | DISCUSSION

We have described a randomization test that can be used for group-level analyses in a within-participant neuroimaging study. This randomization test controls the FA rate under the null hypothesis of statistical independence between the biological data and the explanatory variable. Because the FA rate control of a randomization test does not depend on the test statistic, we discuss ways to design a test statistic that optimizes sensitivity. Specifically, we introduce the min(p) method for combining test statistics with different sensitivity profiles. We performed a set of simulations that demonstrate accurate FA rate control and illustrate the different sensitivity profiles of cluster-based test statistics with different CDTs and cluster definitions. Using the min(p) method for combining these test statistics resulted in a drastic increase of sensitivity, improving on the existing methods for statistical analysis of fMRI data. This increase in sensitivity was not at the expense of the spatial specificity of the inference.

This article belongs to a long tradition of nonparametric statistical methods for the analysis of neuroimaging data (Bullmore et al., 1996; Hayasaka & Nichols, 2003; Maris, 2012; Maris & Oostenveld, 2007; Nichols & Holmes, 2002; Raz, Zheng, Ombao, & Turetsky, 2003; Winkler et al., 2014; Winkler et al., 2016). There are three essential differences between the existing literature and the present paper. First, the present paper builds on the novel null hypothesis of statistical independence between the biological data and the explanatory variable, whereas previous work has mainly focused on exchangeability and distributional symmetry (Maris & Oostenveld, 2007; Nichols & Holmes, 2002; Winkler et al., 2014). Even a proper randomization test for single-participant event-related fMRI data (Raz et al., 2003) was introduced from a perspective that focused on exchangeability. Because the null hypothesis is formulated at the level of the raw data (instead of functions of the raw data, such as regression coefficients), this formal framework allows for a straightforward application to event-related designs. In contrast, null hypotheses about functions of the raw data require distributional assumptions that may be violated (Eklund et al., 2016; Winkler et al., 2014).

Second, the present paper demonstrates the usefulness of the min(p) method for combining test statistics with different sensitivity profiles. The application of this method to the analysis of neuroimaging data is not novel (see Winkler et al., 2016), but the motivation for its application (here, combining different sensitivity profiles) is novel. The idea of constructing test statistics with specific sensitivity profiles is also not novel. In fact, it lies at the heart of the TFCE methodology (Smith & Nichols, 2009). However, it is often unknown which sensitivity profile is optimal for a given dataset, and the min(p) method effectively deals with this ignorance by combining the different sensitivity profiles.

Third, our main quantification of sensitivity involved a measure that indexes coverage probability (identification rate), instead of the usual brain-level hit rate. This quantification is in line with the main scientific interest in current neuroimaging research: identifying the

location of the neural tissue that is affected by some experimental contrast. Especially in terms of coverage probability, our best performing cluster statistic (minp-all) outperformed all the existing statistical methods. In addition, our simulations show that spatial specificity does not appreciably suffer from combining different test statistics (see Figure 10c).

As with every simulation study, its results do not have the status of a mathematical proof. In fact, it cannot be excluded that different results may be obtained with other ingredients for the simulation study (e.g., effect topography, noise correlations, test statistics). In other words, its conclusions depend on the parameters that were manipulated and the ones that were kept constant. An important parameter that was kept constant in our simulation study is the effect topography across our population of participants. Inducing heterogeneity in the participant-specific effect topographies would result in a decrease in the sensitivity of all statistical tests on which we reported. In addition, it would complicate the assessment of the spatial specificity of the statistical inference, both in terms of sensitivity (here, measured by the identification rate) as well as voxel-specific FA rate (here, measured by the false identification rate). Heterogeneity in the effect topography confronts us with the difficulty of defining a group-level effect cluster that represents all the participant-specific effect clusters. This is further complicated by the fact that most fMRI data analyses use spatial smoothing as a part of the preprocessing. The MATLAB code that we have shared is a good starting point for an efficient simulation workflow that can address this issue.

Compared to parametric tests based on GRF theory, our randomization test has the advantage that spatial smoothing is not required. It is also not required to define clusters by thresholding one-sample T-statistics, which have the disadvantage that, via their denominator, they depend on the heterogeneity of the participant-specific effect topographies. When using the randomization distribution as the reference distribution, there is complete freedom with respect to preprocessing as well as the choice of the test statistic. This allows us to make (the type of) spatial smoothing optional and to select from a much broader range of test statistics. As example of this advantage, consider the possibility that, depending on the spatial structure in the data, statistical sensitivity may increase as a result of an appropriate type of spatial smoothing. Depending on the data, different types of spatial smoothing may be optimal; smoothing of the contrast- or T-images, smoothing with or without edge-preservation, and so forth (see Lohmann et al., 2018). Calculating parametric *p*-values for smoothed T-images is a nontrivial statistical problem. In contrast, *p*-values under the randomization distribution can be obtained easily, and they control the FA rate for every type of spatial smoothing. Thus, it is both easy and useful to combine spatial smoothing methods such as LISA (Lohmann et al., 2018) with randomization testing.

Analytic flexibility is one of the main threats of reproducible neuroimaging research (Poldrack et al., 2017). This practice involves that, after collecting the data, the researcher analyses his data in several different ways, with each analysis pipeline typically culminating in a statistical test. Combining the min(p) method with preregistration provides a sensitivity-preserving solution for the FA rate inflation that results from this analytic flexibility. This FA rate inflation follows from two errors: (a) designing analysis pipelines after inspecting the patterns in the data, and (b) failing to correct for multiple testing. The simplest prevention against the first error is preregistration, and the simplest remedy for the second error is Bonferroni correction. However, Bonferroni correction fails to take into account the statistical dependence (correlation) between the different test statistics, and this goes at the expense of statistical sensitivity. The min(p) method is very likely to be a sensitivity-preserving alternative for Bonferroni correction because the randomization distribution of the minimum *p*-value does take this statistical dependence into account.

In conclusion, the present paper describes two statistical innovations for neuroimaging studies: (a) a randomization test for group-level analysis in within-participant studies, and (b) a method for combining test statistics with different sensitivity profiles. These two innovations allow for novel statistical tests that control the FA rate and drastically outperform the existing statistical tests with respect to sensitivity. Future research has to show (a) whether the formal framework used in the present paper can also be used for improving sensitivity in other study types (e.g., studies involving explanatory variables that are not under experimental control), and (b) to what degree the results of our simulation study depend on its specific ingredients.

ACKNOWLEDGMENTS

L. G. was supported by a Veni Grant (451-16-013) from the Netherlands Organization for Scientific Research. Both authors thank Anders Eklund for sharing the preprocessed SPM data, and David Richter for sharing the preprocessed FSL data, and Vladimir Litvak, Andre Marquand, Robert Oostenveld, and Johannes Algermissen for their comments on a previous version of this paper.

CONFLICT OF INTEREST

Both authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Donders Data Repository at https://doi.org/10.34973/zw83-tn77.

ORCID

Eric Maris https://orcid.org/0000-0001-5166-1800

REFERENCES

Bansal, R., & Peterson, B. S. (2018). Cluster-level statistical inference in fMRI datasets: The unexpected behavior of random fields in high dimensions. *Magnetic Resonance Imaging*, *49*, 101–115.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Colcombe, S. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739.

Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., … Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, *35*, 261–277.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLoS One*, *12*(11), e0184923.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage*, *4*(3), 223–235.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220.

Genovese, C. R., Lazar, N. A., & Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*, 870–878.

Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. A. (2018). Effect size and power in fMRI group analysis. *BioRxiv*, 295048.

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

Hayasaka, S., & Nichols, T. E. (2004). Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, *23*, 54–63.

Lindquist, M. A., & Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*, *77*(2), 114–125.

Lohmann, G., Stelzer, J., Lacosse, E., Kumar, V. J., Mueller, K., Kuehn, E., … Scheffler, K. (2018). LISA improves statistical analysis for fMRI. *Nature Communications*, *9*(1), 1–9.

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology*, *49*(4), 549–565. https://doi.org/10.1111/j.1469-8986.2011.01320.x

Maris, E. (2019). Enlarging the scope of randomization and permutation tests in neuroimaging and neuroscience. *BioRxiv*, *685560*. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2019/07/09/685560.full.pdf. https://doi.org/10.1101/685560

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Nichols, T. E., & Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*, 1–25.

Noble, S., Scheinost, D., & Constable, R. T. (2020). Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. *NeuroImage*, *209*, 116468.

Pantazis, D., Nichols, T. E., Baillet, S., & Leahy, R. M. (2005). A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *NeuroImage*, *25*, 383–394.

Pesarin, F. (2001). *Multivariate permutation tests*. New York, NY: Wiley.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115–126. https://doi.org/10.1038/nrn.2016.167

Raz, J., Zheng, H., Ombao, H., & Turetsky, B. (2003). Statistical tests for fMRI based on experimental randomization. *NeuroImage*, *19*(2), 226–232.

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *eLife*, *8*, e47869.

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

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797.

Tippett, L. H. C. (1931). *The methods of statistics*. London, England: Williams & Norgate Ltd.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397.

Winkler, A. M., Webster, M. A., Brooks, J. C., Tracey, I., Smith, S. M., & Nichols, T. E. (2016). Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping*, *37*(4), 1486–1511.

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

Zandbelt, B. (2017). Slice display. *Figshare*. https://doi.org/10.6084/m9.figshare.4742866

How to cite this article: Geerligs L, Maris E. Improving the sensitivity of cluster-based statistics for functional magnetic resonance imaging data. *Hum Brain Mapp*. 2021;42:2746–2765. https://doi.org/10.1002/hbm.25399

APPENDIX

The randomization test for a within-participant design controls the FA rate

Here, we give a formal proof of the fact that the randomization test controls the FA rate. Before the actual proof, we introduce the notation, and give a formal description of the null hypothesis and the randomization test procedure.

Notation

The biological data are denoted by Y, and the explanatory variable by X. The variables Y and X are random variables and their realizations (i.e., the values that were actually observed) are denoted by, respectively, y and x. The biological variable Y is an array of n component data structures Y_r (r = 1, …, n) with realizations y_r, each one corresponding to one participant (indexed by r) that is randomly and independently drawn from some population. In a single-run fMRI experiment, y_r is the multivoxel signal recorded in this run.

The explanatory variable X is a variable of which the relation with the biological variable Y is of scientific interest: stimulus/cue type, task/instruction, and so forth. The variable X is an array of n components X_r (r = 1, …, n) with realizations x_r, each one corresponding to one participant. Every component X_r in turn consists of m subcomponents X_{rs} (s = 1, …, m) with realizations x_{rs}, each one corresponding to one event time. In the following, we will denote X_r as the *condition order*.

The data are depicted in Figure A1. This figure is schematic and applies to both scalar and high-dimensional biological data; the structure of the data arrays is not shown in the figure. Note that it is not possible to separate the participant-specific Y_r in a set of m smaller subcomponent data structures, each one corresponding to one of the

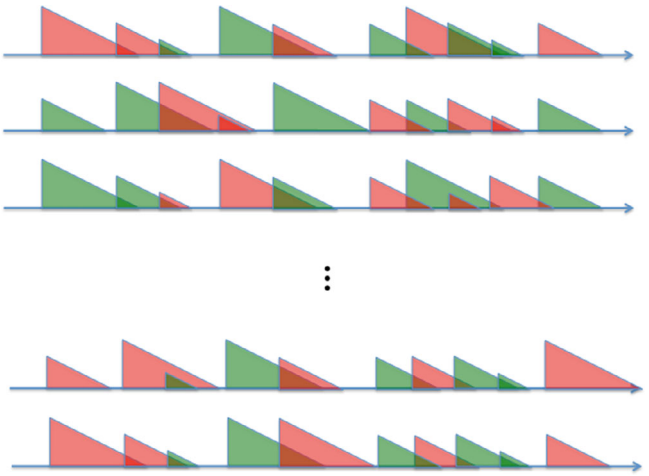


FIGURE A1 Schematic representation of the data of a study with a within-participants manipulation of the explanatory variable. Every timeline (row) corresponds to one participant and every triangle to one event. The colors of the triangles denote the experimental conditions (red = A, green = B), and their heights denote the amplitude of the biological data

m subcomponents X_{rs}. This is because in an MR time series the effects of the different events are superimposed on each other. Also note that there are only two possible condition orders, but that the participants can have different time courses of triangle heights, which correspond to the magnitudes of the event-specific BOLD-responses.

The hypothesis of statistical independence

Our randomization test is a test of the hypothesis of statistical independence between the biological data Y and the explanatory variable X. This hypothesis pertains to the conditional probability distribution of the biological data Y given the explanatory variable X: $f(Y = y|\ X = x)$. The probability distribution of Y may of course depend on other variables besides X, the explanatory variable of interest, but the effect of all these other variables will be considered noise that contributes to the variability of Y for a given realization x of X.

Formally, with our randomization test, we test the null hypothesis of statistical independence between Y and X:

$$f(Y = y|X = x) = f(Y = y), \tag{A1}$$

or, in brief, $f(Y|\ X) = f(Y)$. In the remainder of this article, unless there is a risk for confusion, we will disregard the distinction between a random variable (X, Y) and its realization (x, y). Statistical independence is symmetrical between Y and X, and therefore can also be expressed as follows:

$$f(X|Y) = f(X) \tag{A2}$$

Equation (A2) is useful for proving the FA rate control of the randomization test.

Equations (A1) and (A2) are the most general formulation of the null hypothesis of statistical independence. However, a more specific formulation is possible if the participant-specific component data structures Y_r are statistically independent from each other. In this case, the null hypothesis can be formulated as follows

$$f(Y_r|X_r) = f(Y_r), \text{for } r = 1,\ldots,n \tag{A3}$$

Because the participants are randomly drawn from some population, the functions $f(Y_r|\ X_r)$ and $f(Y_r)$ characterize probability distributions over this population.

The probability distribution $f(X_r)$ is called the randomization distribution, and we assume it to be known, typically because it is under experimental control. The randomization distribution $f(X_r)$, specifies the probabilities of the different condition orders. The proof that will be given in the following applies to all randomization distributions. However, not all randomization distributions are equally interesting from a neurobiological point of view. The interest is almost always in the difference between two experimental conditions *A* and *B*. For a within-participants design, this interest translates into a randomization distribution with nonzero probabilities only for complementary condition orders, such as *ABBA* and *BAAB*. Noncomplementary condition orders (e.g., *ABBA* and *ABAB*) would result in a statistical test that is less sensitive in detecting a difference between *A* and *B*.

The variable Y and its components Y_r denote raw data. Of course, when calculating a test statistic, the raw data will be processed with the goal of extracting the relevant information for some phenomenon of interest (by means of averaging, GLM-based deconvolution, the Fourier transform, …). However, the proof that will be given in the following makes no assumptions that limit this data processing: if the null hypothesis of statistical independence holds for the raw data, it also holds for any function of the raw data. Of course, if this null hypothesis does *not* hold, then the choice of the test statistic may very well affect the probability of rejecting it (i.e., the sensitivity). In general, a well-informed choice of the test statistic (zooming in on the aspect of the data that best reflects the effect) increases the sensitivity.

The randomization test procedure

The randomization test can be performed with an arbitrary test statistic $S(y^{obs}, x^{obs})$, in which y^{obs} and x^{obs} are the realizations of Y and X that were observed in the study. In a study with a within-participants manipulation, the test statistic typically depends on the contrasts between the condition-specific regression coefficients that are obtained from the participant-specific GLM-analyses. These contrasts are combined over the participants, by averaging or by calculating a one-sample (paired-samples) T-statistic.

The reference distribution for the test statistic is obtained by repeatedly calling the same randomization mechanism that also generated x^{obs}, and plugging the resulting random variable X in the test statistic: $S(y^{obs}, X)$. Note that we use the symbol X both to denote the random variable that generates the initial assignment x^{obs}, as well as the random variable that is used to construct the reference

distribution under which the *p*-value is calculated (using the fixed values x^{obs} and y^{obs}). In the following, whenever there is a risk for confusion, we will use X^{rand} to denote the random variable that is used to construct the reference distribution. We will use the same name (randomization distribution) to denote $f(X)$, $f(X^{rand})$, and the reference distribution $f(S(y^{obs}, X^{rand}))$. The *p*-value is calculated by evaluating $S(y^{obs}, x^{obs})$ under $f(S(y^{obs}, X^{rand}))$.

For the implementation of a randomization test, one must know the randomization distribution $f(X)$, which specifies the probabilities of the different condition orders. With two experimental conditions, the most sensitive option is a randomization distribution with two complementary condition orders with equal probabilities. Most researchers want to have an equal number of participants per condition order, and therefore the random assignment involves sampling without replacement. However, the proof of the randomization test's FA rate control does not depend on the details of this randomization mechanism.

To exactly construct the randomization distribution, all possible assignments must be enumerated. When the number of units is large, it is computationally infeasible to perform a complete enumeration. However, in this situation, it is possible to approximate the randomization distribution (with arbitrary accuracy) by randomly drawing values from it. The resulting approximation is denoted as a Monte Carlo estimate, and its accuracy can be quantified by means of a Monte Carlo confidence interval.

The decision about the null hypothesis is taken on the basis of a *p*-value that is obtained under the randomization distribution. For a test statistic of which large values provide evidence against the null hypothesis, the randomization *p*-value can be expressed as $P(S(y^{obs}, X) > S(y^{obs}, x^{obs}))$, in which *P* denotes "probability."

The decision about the null hypothesis (accept or reject) is taken by comparing the randomization *p*-value with the so-called *nominal alpha level*. This nominal alpha level is some a priori value between 0 and 1, typically 0.05 or 0.01. If the randomization *p*-value is less than the nominal alpha level, the null hypothesis is rejected; otherwise, it is accepted.

The randomization test controls the FA rate

We will now prove that, under the null hypothesis of statistical independence, the probability of a randomization test rejecting this null hypothesis is equal to the nominal alpha level. This proof differs from the corresponding proof for a parametric statistical test (e.g., the *t*-test). In the latter case, the test statistic's reference distribution (its probability distribution under the null hypothesis) is known prior to collecting the biological data. In contrast, for a randomization test, the reference distribution depends on y^{obs}. We will deal with this dependence in two steps:

1. We start by proving FA rate control for a specific realization y^{obs} of Y. That is, we will prove *conditional* FA rate control.
2. We prove that conditional FA rate control implies unconditional FA rate control (i.e., independent of y^{obs}).

The randomization test controls the FA rate conditionally given $Y = y^{obs}$

The FA rate is the probability of falsely rejecting the null hypothesis. A false rejection occurs if, under this null hypothesis, the randomization *p*-value $P(S(y^{obs}, X) > S(y^{obs}, x^{obs}))$ is less than the nominal alpha-level (α). The FA rate is evaluated over hypothetical replications of the study, and therefore we must allow for the possibility that the initial assignment (explanatory variable) x^{obs} differs over these replications. We begin by fixing *Y* at y^{obs}, and will therefore consider the conditional FA rate given $Y = y^{obs}$. Now, the randomization *p*-value for a given study is $P(S(y^{obs}, X^{rand}) > S(y^{obs}, X = x^{obs}))$, in which the probability is taken over the realizations of X^{rand}. For given values of y^{obs} and x^{obs}, this *p*-value is a constant, but as a function of the random variable *X*, it is random. Now, the probability of rejecting the null hypothesis equals the probability that this random *p*-value is less than α. In terms of the random test statistic $S(y^{obs}, X)$, this equals the probability that $S(y^{obs}, X)$ is larger than the $(1 - \alpha) \times 100$ percent quantile of the randomization distribution $f(S(y^{obs}, X^{rand}))$. Here, we tacitly assume a one-tailed test. However, our proof generalizes to two-tailed tests in a straightforward way.

Because our objective is to determine the FA rate conditionally given $Y = y^{obs}$, we must know the corresponding conditional probability distribution of $S(y^{obs}, X)$: $f(S(y^{obs}, X) | Y = y^{obs})$. We now make use of the null hypothesis of statistical independence between *X* and *Y*. Specifically, because $S(y^{obs}, X)$ is a function of the random variable *X*, under this null hypothesis, not only *X* but also $S(y^{obs}, X)$ is statistically independent of *Y*. Thus, the conditional probability distribution $f(S(y^{obs}, X) | Y = y^{obs})$ is identical to $f(S(y^{obs}, X))$, which in turn is identical to the randomization distribution $f(S(y^{obs}, X^{rand}))$, whose $(1 - \alpha) \times 100$ percent quantile is used to determine whether the null hypothesis will be rejected. As a consequence, under the null hypothesis, and conditional on $Y = y^{obs}$, the probability that $S(y^{obs}, X)$ is larger than the $(1 - \alpha) \times 100$ percent quantile of the randomization distribution $f(S(y^{obs}, X^{rand}))$ is exactly equal to α. In other words, conditional on $Y = y^{obs}$, the probability of falsely rejecting the null hypothesis is exactly equal to α. This completes our proof of the fact that the randomization controls the FA rate conditionally given $Y = y^{obs}$.

FA rate control conditionally given $Y = y^{obs}$ implies unconditional FA rate control

At first sight, controlling the FA rate in this conditional sense (i.e., conditional on $Y = y^{obs}$) is not very appealing. After all, who is interested in the conditional FA rate given one specific realization of *Y*? However, the FA rate is equal to the critical alpha-level, regardless of whether the *p*-value has a conditional or an unconditional interpretation. This is because, for every realization y^{obs} of *Y* on which we condition, the FA rate is equal to the same critical alpha-level. Therefore, if we average over the probability distribution of the random variable *Y*, the FA rate remains equal to this critical alpha-level. This can also be shown in a short derivation. In this derivation, the FA rate under

the conditional distribution $f(Y, X \mid Y = y^{obs})$ is denoted by $P(\text{Reject}\, H_0 | Y = y^{obs})$, and the FA rate under $f(Y, X)$ by $P(\text{Reject}\, H_0)$. The FA rate $P(\text{Reject}\, H_0)$ is obtained by averaging the conditional FA rate $P(\text{Reject}\, H_0 | Y = y^{obs})$ over the probability distribution $f(Y = y^{obs})$:

$$P(\text{Reject}\, H_0) = \int P(\text{Reject}\, H_0 | Y = y^{obs}) f(Y = y^{obs}) dy^{obs}$$
$$= \alpha \int f(Y = y^{obs}) dy^{obs} = \alpha$$

In the first line of this derivation, we make use of the following equality from elementary probability theory: $P(A) = \int P(A \mid B = b) P(B = b) db$. And in the third line, we make use of the fact that the probability densities $f(Y = y^{obs})$ integrate to 1.

We can conclude that an FA rate that is controlled under the conditional distribution $f(Y, X \mid Y = y^{obs})$ is also controlled under the corresponding unconditional distribution $f(Y, X)$. This conclusion is a special case of the following general fact: for every event (in our case, falsely rejecting the null hypothesis) whose probability is controlled under a conditional distribution, also the probability under the corresponding unconditional distribution is controlled. This general fact will be called the conditioning rationale.

The conditioning rationale is used to prove the unconditional control of the FA or Type 1 error rate, and does not involve a claim about the Type 2 error rate (i.e., the probability that null hypothesis is maintained while in fact the alternative hypothesis is true). This is similar to classical parametric statistics, in which only the Type 1 error rate is controlled. However, different from classical parametric statistics, in the nonparametric framework the researcher is free to choose the test statistic. He may do this on the basis of prior knowledge, with the objective to reduce the Type 2 error rate.