Received: 25 October 2021 | Revised: 31 December 2021 | Accepted: 17 January 2022
DOI: 10.1002/hbm.25795

RESEARCH ARTICLE

WILEY

An integrated cluster-wise significance measure for fMRI analysis

Yunjiang Ge[1] | Gang Chen[2] | James A. Waltz[3] | Liyi Elliot Hong[3] | Peter Kochunov[3] | Shuo Chen[3,4]

[1]Department of Mathematics, University of Maryland-College Park, College Park, Maryland, USA

[2]Scientific and Statistical Computing Core, National Institute of Mental Health, National Institute of Health, Bethesda, Maryland, USA

[3]Maryland Psychiatric Research Center, Department of Psychiatry, School of Medicine, University of Maryland, Catonsville, Maryland, USA

[4]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, Maryland, USA

Correspondence
Shuo Chen, Division of Biostatistics and Bioinformatics, School of Medicine, University of Maryland, 660 W. Redwood Street, Baltimore, MD 21201, USA.
Email: shuochen@som.umaryland.edu

Abstract

Cluster-wise inference is widely used in fMRI analysis. The cluster-level statistic is often obtained by counting the number of intra-cluster voxels which surpass a voxel-level statistical significance threshold. This measure can be sub-optimal regarding the power and false-positive error rate because the suprathreshold voxel count neglects the voxel-wise significance levels and ignores the dependence between voxels. This article aims to provide a new Integrated Cluster-wise significance Measure (ICM) for cluster-level significance determination in cluster-wise fMRI analysis by integrating cluster extent, voxel-level significance (e.g., *p* values), and activation dependence between within-cluster voxels. We develop a computationally efficient strategy for ICM based on probabilistic approximation theories. Consequently, the computational load for ICM-based cluster-wise inference (e.g., permutation tests) is affordable. We validate the proposed method via extensive simulations and then apply it to two fMRI data sets. The results demonstrate that ICM can improve the power with well-controlled family-wise error (FWE).

KEYWORDS
Chernoff bound, cluster, dependent *p*-value, fMRI, spatial correlation

1 | INTRODUCTION

In high-dimensional inference, handling multiple comparison problems remains a popular topic due to its wide applications in scientific fields. Cluster-wise inference is among the most commonly used multiplicity correction approaches for functional magnetic resonance imagining (fMRI) data analysis (Lindquist & Mejia, 2015). This method is a two-step inference procedure including a voxel-level thresholding step to binarize all voxels, and cluster-extent-based inference to decide the cluster-level activation while controlling the family-wise error rate (FWER; Eklund, Nichols, & Knutsson, 2016; Nichols & Holmes, 2002; Poline & Mazoyer, 1993). Generally, inference is used to support the claims of associations between covariates of interest and brain imaging clusters (Bowman, Guo, & Derado, 2007; Lindquist, 2008). In practice, however, the cluster-wise FWER correction approach may lead to inflated FWE (Eklund et al., 2016) due to the violation of model assumptions. To mitigate the inflated error, several adjustment methods have been developed (Eklund, Knutsson, & Nichols, 2019), for example, the parametric voxel-wise inference, and Gaussian random field theory based cluster-wise inference with corrected long residual tail (Cox, Chen, Glen, Reynolds, & Taylor, 2017; Gopinath, Krishnamurthy, Lacey, & Sathian, 2018).

In cluster-wise inference, the selection of both primary threshold and cluster-wise threshold plays a critical role (Ge et al., 2021). In the traditional approach, the cluster-wise threshold is usually decided by the cluster-extent, which is calculated by tallying the number of voxels within the cluster (Lindquist & Mejia, 2015; Zhang, Nichols, & Johnson, 2009). This computationally convenient criterion, however, ignores the variation of p values beyond the predetermined threshold and the dependence structure of voxels within the cluster. Due to the information loss, the cluster-wise inference can be suboptimal regarding the sensitivity and false-positive error rate (Eklund et al., 2016; Woo, Krishnan, & Wager, 2014). To address this challenge, we develop a new approach to calculate the cluster-wise significance by integrating p values of voxels and cluster extent, while accounting for the dependence structure between voxels.

The first step of our proposed method is built on statistical techniques that combine the inference results of multiple hypothesis tests (Heard & Rubin-Delanchy, 2018; Westberg, 1985; Zaykin, Zhivotovsky, Czika, Shao, & Wolfinger, 2007). Various combining methods, such as Fisher's combined probability test (Fisher, 1992), Stouffer's statistic (Stouffer, 1949), Tippett's method (Tippett, 1931), and recent approaches with a Cauchy distribution (Liu & Xie, 2020) or a harmonic mean p-value (Wilson, 2019), have been developed and commonly used due to their good properties on consistency and accuracy (Alves & Yu, 2014; Brown, 1975; Chen et al., 2014; Liu & Xie, 2020). Among the combining methods, Fisher's method is the earliest and the most popular one. It allows both independent and dependent multiple tests because the summation of the log-transformed p values can be established by an asymptotic Chi-sqaure distribution (Hayasaka & Nichols, 2004; Lazar, Luna, Sweeney, & Eddy, 2002; Winkler et al., 2016; Zhang et al., 2009), and numerical methods with high accuracy are available to approximate its parameters (Brown, 1975; Kost & McDermott, 2002).

In practice, combining methods are not directly applicable to our application due to two major limitations. First, the dependence structure in brain imaging data is spatially constrained (Derado, Bowman, & Kilts, 2010). The existing combining methods do not fully address the spatial dependence in brain imaging data. Even though the permutation framework maintains the spatial structure and allows the unknown distribution of the combining methods, (Hayasaka & Nichols, 2004; Lindquist & Mejia, 2015) the accuracy of the combined statistics that rely on parametric distributions for a given cluster can be compromised if the correlations between voxels are high (Efron, 2007). The biased estimation can make it hard to distinguish the noise and the signal of interest (Leek & Storey, 2008). A second major limitation is that, in cluster-wise inference, voxel-wise p values are restricted by their floating-point representation; that is voxel-wise p values are often thresholded by a small value no greater than .001 (Woo et al., 2014). Therefore, the cluster-wise p values can be extremely small ($<10^{-100}$) and thus we are unable to distinguish two clusters with similarly combined p values but carrying much different information (e.g., sizes, voxel-wise p values, spatial dependence). This is incompatible with the permutation test for controlling family-wise error (FWE) and limits the overall utility.

To address the issues of information loss and computation compatibility, we developed a tailored method to incorporate the spatially constrained dependence structure between voxels into combining dependent p values of intra-cluster voxels. Additionally, we propose a new strategy to substitute the exact cluster-wise p values with computationally efficient probability bound. We further prove the log-transformed probability bound has a monotonic relationship with the exact cluster-wise p-value, and therefore can conveniently be adopted by the permutation tests to accurately rank the significance levels of clusters. Our method is also compatible with the Threshold-Free Cluster Enhancement (TFCE) method (Smith & Nichols, 2009) because our enriched significant level is a suitable substitute for the cluster extent. Therefore, it can become a general tool for cluster-wise analysis.

We further organize the paper as follows. In Section 2, we introduce our method with technical details. We provide a task-based fMRI data example and a resting-state fMRI data example to demonstrate the effectiveness of our statistic in Section 3. We then evaluate the performance of our method through simulation studies in Section 4. In the last section, we discuss and summarize the new approach. Additional proofs, parameter derivations, and simulation results are provided in the Appendix.

2 | METHODS

2.1 | Background

In fMRI analysis, our interest is to investigate the association between clinical or experimental covariates and localized brain activation or connectivity (to seed voxels). Conventionally, statistical analysis is conducted on each voxel in the whole brain or in specified spaces. In a simple case, it can be written in a general linear regression model (GLM) framework. Consider a sample of $i = 1,\ldots, N$ subjects. There are $j = 1,\ldots, V$ voxels in an fMRI scan for one subject, and each voxel contains the localized measurement of brain activity as the outcome $\boldsymbol{Y}_{\mathrm{N}\times\mathrm{V}}$. The subject-level covariates are denoted by $\boldsymbol{X}_{\mathrm{N}\times\mathrm{q}}$, where q represents the covariates, such as clinical status or demographic variables. The corresponding parameters are given by $\boldsymbol{\beta}_{\mathrm{q}\times\mathrm{V}}$. Consider one specific covariate of interest. For i-th subject, $\boldsymbol{Y}_{\mathbf{i}} = X_i\boldsymbol{\beta} + \epsilon_{\mathbf{i}}$, $i = 1,\ldots, N$, where $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_{\mathrm{V}})'$ and the error term is $\epsilon_{\mathbf{i}} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. We wish to simultaneously test the hypothesis

$$H_{0j}:\beta_j = 0 \quad \text{versus} \quad H_{Aj}:\beta_j \neq 0$$

to search for the voxels that are correlated with the task/behavior. Let $\boldsymbol{X}_{\mathrm{N}\times 1}$ be the covariate of our primary interest, the corresponding estimated parameter $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_{\mathrm{V}})'$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The voxel-wise test statistics are denoted by $\boldsymbol{T} = \{t_1,\ldots, t_{\mathrm{V}}\}$ and $\boldsymbol{P} = \{p_1,\ldots, p_{\mathrm{V}}\}$, respectively.

Based on the voxel-level test statistics, the two-step cluster-wise inference further extracts findings at the cluster-level, which gains additional power (Nichols & Hayasaka, 2003). The primary thresholding step is considered to be a screening step. With a given threshold p_θ (e.g., $p_\theta < .001$), the voxels can be written into the sets $\boldsymbol{V} = \boldsymbol{V_0} \cup \boldsymbol{V_A}$, where $\boldsymbol{V_0} = \{v \in \boldsymbol{V}: p_v > p_\theta\}$ and $\boldsymbol{V_A} = \boldsymbol{V} \backslash \boldsymbol{V_0}$. At cluster level, the inference is conducted on the clusters that are consisted of contiguous voxels in $\boldsymbol{V}_{\mathbf{A}}$. Specifically, in a 3D volume, we denote the clusters by $c = 1, \ldots, C$, where each cluster c is a set of voxels satisfying $\boldsymbol{V}_{\mathbf{c}} = \{v_j \in \boldsymbol{V}_{\mathbf{A}}$: all voxels in the cluster c with a neighbor$\}$. Denote the cardinality measure of $\boldsymbol{V}_{\mathbf{c}}$ as $n_c := |\boldsymbol{V}_{\mathbf{c}}|$. We remove the singletons from $\boldsymbol{V}_{\mathbf{A}}$ to obtain all the clusters in $\boldsymbol{V}$, which is given by $\boldsymbol{V}_{\boldsymbol{A}}^{*} = \boldsymbol{V}_{\mathbf{1}} \cup \ldots \cup \boldsymbol{V}_{\boldsymbol{C}}$. We further conduct inference on the subsets in $\boldsymbol{V}_{\mathbf{A}}^{*}$.

Each cluster c contains n_c voxels with their test statistics $t_j^{(c)}, j = 1, \ldots, n_c$, significance levels $p_j^{(c)}$ and dependence structure $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{(c)}$. Our ICM utilizes all of above information and provides a computationally convenient significance level. The general procedure is described in Figure 1.

In particular, each piece of the information is described as below:

1. *Cluster-extent*: This is the cardinality measure of the set $\boldsymbol{V}_{\mathbf{c}}$, which is given by n_c.
2. *Voxel-wise Statistical Significance Levels*: For cluster c, voxels within $\boldsymbol{V}_{\mathbf{c}}$ have p values $\mathbf{P}_{\mathbf{c}} = \left\{p_1^{(c)}, \ldots, p_{n_c}^{(c)}\right\}$. We use the significant level of each voxel to represent the association strength between a voxel and the regressor of interest.
3. *Dependence Structure*: The covariance structure of estimated parameter $\hat{\boldsymbol{\beta}}$ in cluster c is denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{(c)}$.

The cluster-extent and the association strength are widely used statistics in the random field theory based methods, which can efficiently accommodate either focal or spatially extended signals (Poline, Worsley, Evans, & Friston, 1997; Worsley, Evans, Marrett, & Neelin, 1992; Zhang et al., 2009). Here, we provide the cluster-level statistic that incorporates the spatial dependence in addition to the cluster-extent and the association strength from all voxels within the cluster.

2.2 | Cluster-wise statistic combining *p* values of voxels with dependence

2.2.1 | A probabilistic model of combining dependent *p* values

Given a cluster c with n_c voxels, we have a set of dependent p values $\boldsymbol{P}_c = \{p_1, \cdots, p_{n_c}\}$. When voxels are *independent*, Fisher's method can be used to compute the sum of log-transformed p values: $\Psi_c = -2\sum_{j=1}^{n_c} \log p_j \sim \chi^2_{2n_c}$. However, $\boldsymbol{P}_c = \{p_1, \cdots, p_{n_c}\}$ are dependent in our application because voxels are correlated in a cluster. Therefore, $\Psi_c = -2\sum_{j=1}^{n_c} \log p_j$ follows a scaled Chi-square distribution $a\chi^2_f$, where a is the scale parameter and f denotes the the degree of freedom (*df*; Brown, 1975).

Thus a and f can be calculated based on the first two moments of Ψ_c and $a\chi^2_f$. Specifically, we have

$$a = Var(\Psi_c)/[2E(\Psi_c)], f = 2\{E(\Psi_c)\}^2/Var(\Psi_c).$$

In the above formula, $E(\Psi_c) = 2n_c$ is determined by the cluster extent, while $Var(\Psi_c) = \sum_{j,k} cov(-2\log p_j, -2\log p_k)$ and each entry $cov(-2\log p_j, -2\log p_k)$ can be calculated based on correlation between j-th and k-th voxels $r_{jk} = corr\left(\hat{\beta}_j, \hat{\beta}_k\right)$ (Kost & McDermott, 2002; Krylov & Stroud, 2006). The r_{jk} can either be calculated empirically from the data sample or approximated by a parametric model with spatial information (e.g., Matérn correlation). The detailed procedure for estimating r_{jk} is provided in the Appendix A.

We denote $\boldsymbol{R}_{\mathrm{c}}$ as the correlation matrix for the estimated parameters of interest $\hat{\boldsymbol{\beta}}$ in cluster c. Thus, $\boldsymbol{R}_c = \boldsymbol{D}_c^{-1/2} \Sigma_{\beta}^{(c)} \boldsymbol{D}_c^{-1/2}$ where $\boldsymbol{D}_{\mathrm{c}} = \mathrm{diag}(\Sigma_{\boldsymbol{\beta}}{}^{(c)})$. Let $\boldsymbol{U}_{\mathrm{c}}$ be approximated by

$$\boldsymbol{U}_c \approx 4\boldsymbol{I}_c + 3.263\boldsymbol{L}_c + 0.71\boldsymbol{L}_c^{\circ 2} + 0.027\boldsymbol{L}_c^{\circ 3}, \tag{1}$$

where $\boldsymbol{L}_{\mathrm{c}} = \boldsymbol{R}_{\mathrm{c}} - \boldsymbol{I}_{\mathrm{c}}$, $\boldsymbol{I}_{\mathrm{c}}$ is the identity matrix of size n_{c}, and "$\circ$" is Hadamard product. Numerically, we have $Var(\Psi_c) \approx \mathbf{1}_{\mathbf{c}}{}^{\mathrm{T}}\boldsymbol{U}_{\mathbf{c}}\mathbf{1}_{\mathbf{c}}$, where $\mathbf{1}_{\mathbf{c}}$ is a $n_c \times 1$ vector of ones, and the coefficients are approximated by polynomial regression models with accuracy 10^{-4}. (Brown, 1975; Kost & McDermott, 2002).

We further let the cluster-wise statistic of dependent p values be $T_c = \frac{\Psi_c}{a}$, and the corresponding p-value for the cluster c is

$$p_c = 1 - \varphi_f(T_c) \tag{2}$$

where φ_f is the CDF of χ^2_f. The cluster-wise p-value p_c in Equation (2) integrates the three important features of a cluster because (a) the test statistic Ψ_c is the sum of voxel-level significance $\boldsymbol{P}_c = \{p_1, \cdots, p_{n_c}\}$;

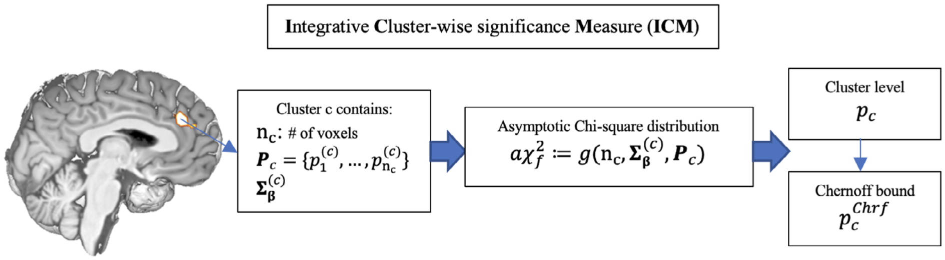

Integrative Cluster-wise significance Measure (ICM)
Cluster c contains:
$\mathrm{n_c}$: # of voxels
$\boldsymbol{P}_c = \{p_1^{(c)}, \ldots, p_{\mathrm{n_c}}^{(c)}\}$
$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(c)}$
Asymptotic Chi-square distribution
$a\chi_f^2 := g(\mathrm{n_c}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(c)}, \boldsymbol{P}_c)$
Cluster level
p_c
Chernoff bound
p_c^{Chrf}

FIGURE 1 An overview of ICM: we first integrate n_c, P_c, $\Sigma_\beta{}^{(c)}$ for cluster c to compute a statistic $a\chi_f^2$ asymptotically following a scaled Chi-square distribution. We use the corresponding p-value P_c instead of the df-based test statistic because p values can be compared between clusters with different sizes. Lastly, the log-transformable Chernoff bound for P_c is derived to overcome the floating-point limitation for extremely small values of P_c, which are common for fMRI clusters

and (b) the scale parameter and *df* of the reference χ^2-distribution reflect a combination of the cluster extent n_c and covariance between voxels $\mathbf{\Sigma}^{(c)}$.

2.2.2 | An approximate bound for cluster-wise p_c

In practice, the cluster-wise *p*-value p_c can be extremely small (e.g., $p_c < 10^{-100}$) for a commonly observed cluster of hundreds of voxels. However, the cumulative probability function in most software yields a cluster-wise *p*-value equal to 0 regardless of the true p_c. This may cause a serious issue for the following multiple testing correction when controlling the family-wise error rate. For example, in a permutation test, cluster-wise *p* values for two cluster *c* and *c′* are both 0 based on the software cumulative probability function, although $p_c \neq p_{c'}$. The indistinguishable *p* values in permutation iterations can prohibit the proper inference of permutation test, and thus limit the practical utilization of the proposed cluster-wise *p*-value method. To address this issue, we use a computationally efficient probabilistic bound as a valid approximate for the exact p_c and provide its monotonic property with p_c.

In the light of p_c being a tail-end area of the χ^2_f distribution, we apply the Chernoff bound, which is an exponentially decreasing power-law bound on tail-distributions to approximate p_c (Chernoff, 1952). Obtaining from the moment generating function of χ^2_f, the probabilistic bound of $\varphi_f(T_c)$ is given by $P\left(\chi^2_f \geq T_c\right) \leq \left(\frac{T_c}{f}\right)^{f/2} \exp\left(\frac{f-T_c}{2}\right)$. The RHS of the inequality is the Chernoff upper bound for the significance level p_c, denoted by p_c^{Chrf}. Thus, our integrated cluster-wise significance measure (ICM) can be calculated by

$$p_c^{Chrf} = \left(\frac{T_c}{f}\right)^{\frac{f}{2}} \exp\left(\frac{f-T_c}{2}\right) \tag{3}$$

where the calculation details of *f*, T_c are given in the last section (Brown, 1975; Kost & McDermott, 2002). Thus, Equation (3) is a closed-form approximate for the p_c. We have the Lemma 1 (proof in Appendix B.1) to ensure p_c^{Chrf} is a monotonic function of p_c, and thus can substitute p_c by log-transformed p_c^{Chrf} in the permutation test.

> **Lemma 1.** *Let two clusters c_1 and c_2 have approximated df satisfying $f_{c_1} = f_{c_2}$. If $p_{c_1} < p_{c_2}$, the Chernoff bound for p_{c_1}, p_{c_2} satisfy $p_{c_1}^{Chrf} < p_{c_2}^{Chrf}$.*

2.2.3 | An integrated measure for cluster-wise significance

The proposed ICM p_c^{Chrf} integrates the information of cluster-extent, voxel-wise statistical significance levels, and spatial dependence between voxels within the cluster. In particular, we let $\rho_c = \mathbf{1_c}^{\mathrm{T}}(\mathbf{U_c} - 4\mathbf{I_c})\mathbf{1_c}$, $\mathbf{U_c}$ referring to Equation (1), which is the summation of the off-diagonal elements in the correlation matrix. Note that we use ρ_c to represent the spatial dependence so that we can exclude the cluster-extent effect.

We re-write the ICM p_c^{Chrf} for cluster *c* as a function of n_c, $\mathbf{P_c}$ and ρ_c as follows.

$$\text{Cluster extent}: g_1(n_c, \mathbf{P_c}, \rho_c) = n_c$$

$$\text{Voxel-wise Statistical Significance Levels}: g_2(n_c, \mathbf{P_c}, \rho_c) = -2\sum_{j=1}^{n_c} \log p_j$$

$$\text{Spatial Dependence}: g_3(n_c, \mathbf{P_c}, \rho_c) = \rho_c$$

Based on $\boldsymbol{g}(n_c, \mathbf{P_c}, \rho_c) = (g_1(n_c, \mathbf{P_c}, \rho_c), g_2(n_c, \mathbf{P_c}, \rho_c), g_3(n_c, \mathbf{P_c}, \rho_c))$, we have the composite function

$$p_c^{Chrf} = (p^\circ \boldsymbol{g})(n_c, \mathbf{P_c}, \rho_c) = \left(\frac{-2\sum_{j=1}^{n_c} \log p_j}{2n_c}\right)^{\frac{4n_c^2}{4n_c+\rho_c}} \exp\frac{2n_c\left(2n_c + 2\sum_{j=1}^{n_c} \log p_j\right)}{4n_c+\rho_c} \tag{4}$$

where "∘" is the function composition operator and $p(\boldsymbol{g}) = \left(\frac{g_2}{2g_1}\right)^{\frac{4g_1^2}{4g_1+g_3}} \exp\frac{2g_1(2g_1-g_2)}{4g_1+g_3}$. Through this derivation, the calculation of p_c^{Cherf} is fairly straightforward and can be conveniently implemented in various software packages. As the p_c^{Cherf} has a complicated form, we explore some noticeable properties by its statistical parameters in the next section.

2.3 | Properties of ICM

The ICM is jointly decided by the cluster-extent, voxel-wise statistical significance levels, and spatial dependence. In this section, we specifically explore the relationships between ICM and these three factors, and thus better understand their joint influence on p_c^{Cherf} instead of n_c alone in classical cluster-wise inference. Based on Lemma 1, we have Theorem 1 showing that a smaller average intra-cluster voxel-wise *p*-value can lead to a lower p_c^{Cherf} value. Theorem 2 states that a larger *df* results in a more significant p_c^{Cherf} value. Proposition 1 adds a more restricted condition on Theorem 2, which concludes that the higher dependence leads to a less signficant p_c^{Cherf} value. All proofs are provided in the Appendix B.

We firstly provide a visualized demonstration of p_c^{Chrf} on a 3D surface in Figure 2. The average *p*-value is the mean of voxel-wise significance levels, while we use the average correlation to represent the general dependence level of a cluster. The scale is adjusted to logarithm base 10 for the convenience.

To begin with, we consider a fixed *df* and evaluate the effect of average intra-cluster voxel-wise *p*-value, $s_c = \frac{\Psi_c}{n_c}$ on ICM. With a given *df*, we have Theorem 1 as follows.

> **Theorem 1.** Let two clusters c_1 and c_2 have approximated df satisfying $f_{c_1} = f_{c_2}$. For their average signal strength, if $s_{c_1} < s_{c_2}$, the Chernoff bound for p_{c_1}, p_{c_2} satisfy $p_{c_1}^{Chrf} > p_{c_2}^{Chrf}$.

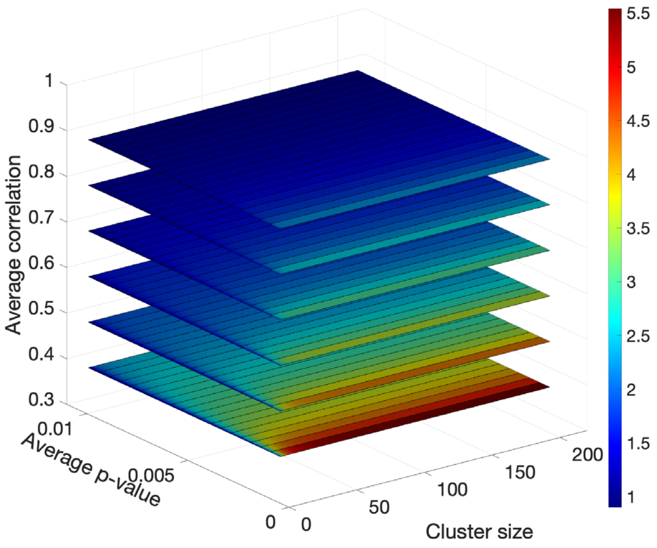


FIGURE 2 Relationship between p_c^{Chrf} and cluster extent (x-axis), average p-value (y-axis), average correlation within-cluster (z-axis). Cluster extent ranges from 1 to 200; average p-value ranges from 0.0005 to 0.01; average correlation within-cluster ranges from 0.4 to 0.9. p_c^{Chrf} scale is adjusted to logarithm base 10

We further use a simulation data example to demonstrate this property in Appendix C.

Since df is jointly determined by the cluster extent and dependence, we evaluate this joint effect with a fixed s_c value.

Theorem 2. Let two clusters c_1 and c_2 have same average signal strength $s_{c_1} = s_{c_2}$. If $f_{c_1} < f_{c2}$, then the Chernoff bound for p_{c_1}, p_{c_2} satisfy $p_{c_1}^{Chrf} > p_{c_2}^{Chrf}$.

Under fixed s_c, the p_c^{Chrf} tends to be smaller with larger approximated df.

If we further restrict two clusters to have the same cluster size based on Theorem 2 condition (i.e., fixed s_c), we can conclude Proposition 1 on the total dependence.

Proposition 1. *Given the ratio* $s_c = \frac{\Psi_c}{n_c}$ *and cluster size* n_c *fixed, the* p_c^{Chrf} *is increasing if the total dependence* ρ_c *is increasing.*

The proposition explains the phenomenon that the higher smoothness level leads to a less significant p_c^{Chrf} value.

In summary, the ICM p_c^{Chrf} is more significant for a cluster with stronger average signal, larger cluster extent, and less dependence between voxels. In the data example and simulations, we show that ICM can outperform conventional cluster-wise inference methods.

3 | DATA EXAMPLES

To provide real-data demonstrations of the ICM, we performed cluster-wise inference on both task-based and resting-state fMRI(rs-fMRI) data sets. The task-based study involved the collection of brain and behavioral data related to reinforcement learning, and we aim to evaluate ICM and the cluster-extent method for sensitivity and reproducibility. We also conduct seed-voxel based Function Connectivity (rsFC) analysis on rs-fMRI data to explore the FWE-control performance of ICM.

3.1 | Dataset 1: Reinforcement learning task-based fMRI study

We apply ICM to a full task-based fMRI dataset (with all participants included) as well as a sub-sample determined through random selection. We intend to see (a) if ICM can potentially detect more biologically meaningful regions, and (b) if the detected regions from full sample and sub-sample are consistent.

3.1.1 | Data preparation

Task-based fMRI data on reinforcement learning (RL) were collected from 26 schizophrenia patients (SZ) and 26 healthy volunteers (HV) at the University of Maryland Center for Brain Imaging Research. Nineteen participants were female. The average age of all participants was 36 ± 12, with no difference between gender groups ($p = .71$) or patient-control groups ($p = .68$). The participants learned three probabilistic discriminations, including the potential gain (gain/miss [GM]), nonmonetary (correct/incorrect [CI]), and potential loss (loss/avoid [LA]). Participants performed 240 trials over the course of four runs of 60 trials, and the functional MRI data were acquired simultaneously with task performance.

A 3T Siemens Trio scanner (Erlangen, Germany) was used to measure $T2^*$-weighted blood oxygen level-dependent (BOLD) effects with the following parameters: 81 2-mm axial slices, 128 × 128 matrix, FOV = 22 × 22 cm, TR = 2 s, 1.5 × 1.5 × 1.5 mm voxel size. A whole-brain T1-weighted structural image was also acquired in each session for anatomical reference. Voxel time series were normalized with the AFNI software package. The subject-specific beta coefficients were obtained from two sets of regression analyses. One set contains binary regressors corresponding to three probabilistic discriminations (GM, CI, and LA) and two possible outcomes (gain/neutral, neutral/loss, and correct/incorrect); the other set has parametric regressors that are derived from the results of individual behavior for estimation of reward prediction errors (RPEs), which signal mismatches between expected and obtained outcomes. In addition, head-motion vectors were included in each regression model as regressors of no interest (Waltz et al., 2018).

3.1.2 | Data analysis

The Aberrant Salience Inventory (ASI) is a measure of unusual experiences of salience in the environment, as well as general psychosis proneness among clinical and nonclinical participants (Cicero, Kerns, & McCarthy, 2010). We perform a voxel-wise regression analysis across all participants with ASI total score as our primary regressor of interest, and age, gender, group, and educational level as nuisance regressors. We aimed to identify regions whose activity modulates the relationship between psychosis proneness and evoked responses to RPEs. For more rigorous FWE-control, we applied a voxel-level primary threshold of $p < .0001$, before performing permutation testing (controlled at FWE < 0.05) on supra-threshold voxels to generate the cluster-extent threshold and ICM threshold. We also applied other combining p methods that do not specifically account for the spatial dependence, such as combined p approximated by Cauchy distribution (denoted by p^{Cauchy}), and uncorrected Fisher's combined p (denoted by p^{Fisher}).

To validate the reproducibility of ICM, we randomly sampled 36 subjects from the full sample (18 SZs and 18 HCs). The primary regressor of interest and adjusted covariates remain the same, as well as the primary threshold $p < .0001$ and the entire cluster-wise inference procedure. We aim to compare the findings from the sub-sample with those from the full sample.

3.1.3 | Results

We identified associations between individual ASI scores and neural responses evoked by RPEs in two regions: (1) right middle temporal gyrus and (2) right inferior temporal gyrus. Cluster (1) had 238 voxels, and the peak voxel is at (64, −50, −4). Cluster (2) had 116 voxels, and the peak voxel is at (51, −18, −37). Detailed information of the clusters are summarized in Table 1. A demonstration of the above regions on a 3D surface model (Kochunov et al., 2001; Lancaster et al., 2010, 2012) is given in Figure 3.

We then compared the findings using ICM threshold with those using the cluster-extent threshold and other combining p methods p^{Cauchy} and p^{Fisher}, controlling the FWE at $\alpha = .05$. The cluster-extent threshold is 135, which excludes the cluster (2) at inferior temporal gyrus. The p^{Cauchy} and p^{Fisher} thresholds tend to be so stringent that no cluster can pass their threshold.

In the sub-sample, the ICM detects one activated region with cluster size 167, locating at right inferior temporal gyrus. Refer to Table 1 and Figure 4 for details of the finding. This cluster overlapped to a large degree with the Cluster (2) in the full sample. We computed the Jaccard index of this activated cluster in the sub-sample and Cluster (2) in the full sample at 0.44. No cluster passes the cluster-extent threshold in this sub-sample.

3.1.4 | Remarks

We further explored the biological features of the regions detected in the full sample and sub-sample. The two regions discovered in the full sample are involved in various cognitive processes, including the multi-modal sensory integration on two regions together, language and semantic memory processing on middle temporal gyrus, and visual perception on inferior temporal gyrus (Cabeza & Nyberg, 2000; Chao, Haxby, & Martin, 1999; Herath, Kinomura, & Roland, 2001; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999; Mesulam, 1998; Onitsuka et al., 2004; Tranel, Damasio, & Damasio, 1997). It has been established that patients with a history of psychosis have elevated ASI (Cicero et al., 2010; Raballo et al., 2019). Similar studies on functional deficits that are associated with abnormal RPE are often reported among schizophrenia patients in those regions (Boehme et al., 2015; Murray et al., 2008; Roiser, Howes, Chaddock, Joyce, & McGuire, 2013; Takemura, Samejima, Vogels, Sakagami, & Okuda, 2011).

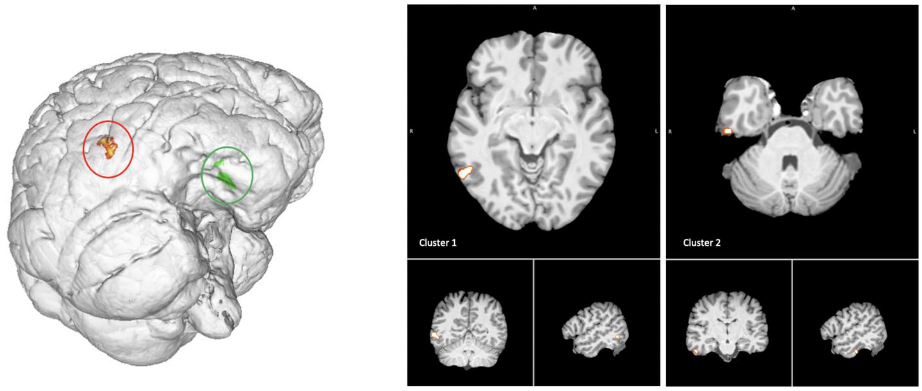
3.2 | Dataset 2: rs-fMRI data with noninformative covairates

In this experiment, our goal is to empirically measure the capability of controlling FWER when no true signal presents. The rsFC analysis was performed, where a 10 mm spherical seed was placed centering on the posterior cingulate cortex (PCC) at (−5, −49, 40), and the correlations were calculated and normalized (with Fisher's Z transformation) between the rest of voxels and the seed. The treatment was chlorpromazine (CPZ) equivalent daily dose (Ge et al., 2021; Hare et al., 2017, 2021). We randomly shuffled the treatment covariate to generate false-positive clusters.

TABLE 1 Significant clusters information detected by ICM: (a) when sample size is full ($n = 52$) and (b) when sample size is two third of the full sample ($n = 36$)

Sample size	Clusters	Size	MNI: Peak voxel (*x*, *y*, *z*)	BA	Label	Function
$n = 52$	Cluster 1	238	65, −51, −7	R21	Right middle temporal gyrus	Cluster 1&2 together subserve language and semantic memory
	Cluster 2	116	51, −17, −45	R20	Right inferior temporal gyrus	Processing, visual perception, and multimodal sensory integration.
$n = 36$	Cluster 1	167	62, −59, −8	R37&R20	Right inferior temporal gyrus	Processes visual stimuli and memory recall

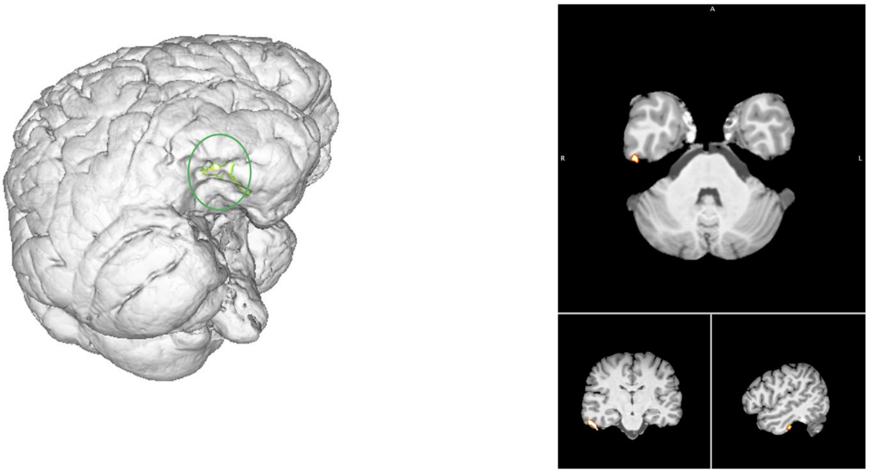
Abbreviations: BA, Brodmann Area; MNI, Montreal Neurological Institute; R, right.



(a) Activated regions on 3D surface

(b) Activated regions projected on 2D images

FIGURE 3 Panels (a) and (b) together show the two activated regions discovered by ICM. On 3D surface model in (a), the regions are displayed and circled out with different colors. In (b), three views of each cluster are displayed, respectively



(a) Activated region on 3D surface

(b) Activated regions projected on 2D images

FIGURE 4 Panels (a) and (b) together show the activated regions discovered by ICM. On 3D surface model in (a), the region is displayed and circled out with green. This finding matches the Cluster 2 in Figure 3

3.2.1 | Data preparation

We collected resting-state fMRI (Rs-fMRI) data of 92 schizophrenia patients (SZs) at University of Maryland Center for Brain Imaging Research. The average age of the SZ cohort was 35.5 ± 13.2, including 26 females. The T2*-weighted BOLD effects were measured by a Siemens 3T TRIO MRI (Erlangen, Germany) system equipped with a 32-channel phase array head coil. The imaging parameters were given as follows: $\mathrm{TR} = 2\,\mathrm{s}$, $\mathrm{TE} = 30\,\mathrm{ms}$, flip angle $= 90^{\circ}$, $\mathrm{FOV} = 22 \times 22$ mm, 128×128 matrix, $3 \times 3 \times 3\ mm$ voxel size.

Rs-fMRI data was preprocessed with the Data Processing & Analysis for (resting-state) Brain Imaging (DPABI) toolbox (Yan, Wang, Zuo, & Zang, 2016). The raw data underwent motion correction, slice-timing correction, and normalization to the MNI space. Regression models on motion parameters and physiological signals were also applied to ensure the spurious motion and physiological artifacts did not drive observed effects in the statistical analyses. Images were smoothed with an 8 mm FWHM Gaussian kernel. Framewise displacement was calculated for each image to differentiate head realignment parameters, which generates a six-dimensional time series to represents the head motion (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012). All individuals have mean framewise displacement <0.25 to control the potential confounding from the motion artifacts.

3.2.2 | Data analysis

In our current study, we only randomly shuffled the CPZ dose values and kept other variables unchanged. Through this step, any activation would be considered as false positive. We conducted study with primary threshold $p < .005$, $p < .001$, $p < .0005$, and performed permutation test with FWE controlled at 5%.

3.2.3 | Results

Among 100 random samples, the ICM and p^{Cauchy} yielded to a well-controlled FWE while the cluster-extent threshold resulted in a much inflated FWER (Table 2). The FWE-control for p^{Fisher} is not applicable due to the computational restriction of the software. Specifically, the minimal cluster-wise p^{Fisher} p values for most permutation iterations are 0 and thus not distinguishable from each other.

TABLE 2 Performance of FWE-control

	Primary threshold		
	$p < .005$	$p < .001$	$p < .0005$
FWER			
ICM	6% (1%, 11%)	7% (2%, 12%)	5% (1%, 9%)
Cluster extent	20% (12%, 28%)	22% (14%, 30%)	20% (12%, 28%)
p^{Fisher}	NA	NA	NA
p^{Cauchy}	5% (1%, 11%)	7% (2%, 12%)	7% (2%, 12%)

The ICM identifies consistent significant clusters from a full sample to its sub-sample. It also strictly controls FWER at its given levels (e.g., 5%). Therefore, our ICM provides a convenient alternative with improved sensitivity and well-controlled false positive rate.

4 | SIMULATION

To evaluate the performance of the proposed ICM, we conduct simulation studies with different types of imaging patterns. We further test ICM on its ability of excluding the FWER. We also evaluate the effect of primary threshold on the ICM.

We apply a commonly-used two-group (i.e., cases vs. controls) scenario, which can be easily extended to the regression setting. We generate two-dimensional images contain $100 \times 100 = 10{,}000$ voxels with a common setting that the voxels from null set follow a normal distribution $N(0, 1)$ in both groups, while the non-null voxels in case group follow a normal distribution $N(\mu, 1)$. The signal-to-noise ratio (SNR) is the reciprocal of the coefficient of variation, given by $SNR = \mu/\sigma$, where the $\sigma = 1$ allows the difference of group means to be the true positive effect size (ES) which is equivalent to Cohen's d. All the images are smoothed with a Gaussian filter, with FWHM equivalent to 4, 6, or 8 mm. These smoothness levels simulate the popular smoothing kernels in the real fMRI data. We further let the number of subjects per group be 30, 60, and 100 to test the performance on different sample sizes. Major criteria are voxel-wise true positive rate (TPR), true discovery rate (TDR), and cluster-wise FWER.

4.1 | General performance under common distribution patterns

We first test the general performance of ICM on the images with the commonly seen distribution patterns.

4.1.1 | Pattern I

The underlying truth for Pattern I contains a squared area $N_0 = 7 \times 7 = 49$ voxels in the center, see Figure 5a left. Images are smoothed with FWHM = 4 mm, ES = 0.4, 0.6, and 0.8; FWHM = 6 or 8 mm, ES = 0.2, 0.4, and 0.6. We compare the results with cluster-extent threshold results. The primary threshold is $p < .001$. In this pattern, we calculate the voxel-wise TPR and TDR, and cluster-wise FWER.

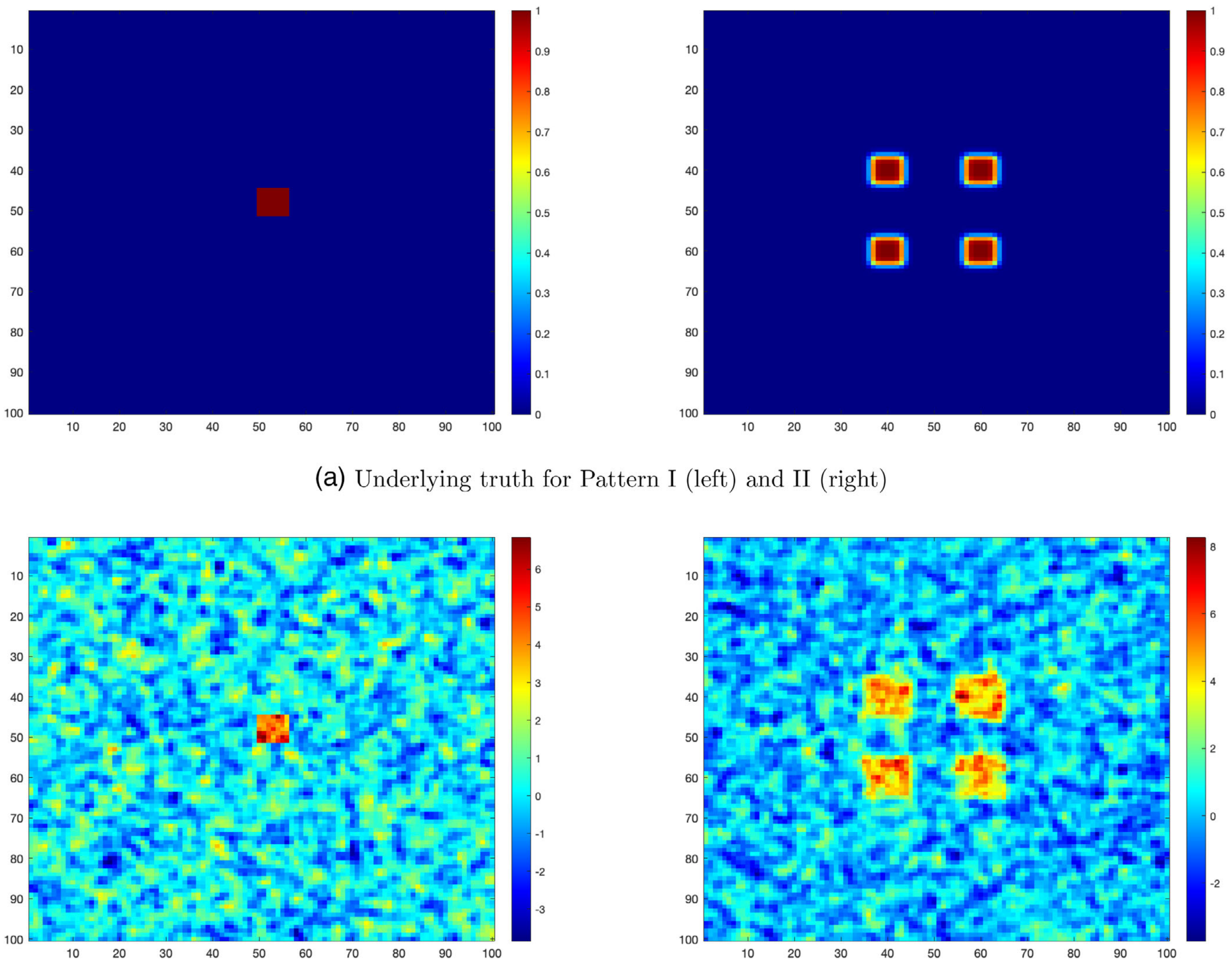
4.1.2 | Pattern II

The Pattern II has irregular-shaped underlying truth. In a three-dimensional brain space, the sensitivity (TPR) and the precision (TDR) are both important since the false positive voxels within a significant cluster can extend the detected region to multiple brain areas that are correlated with different functions. The underlying truth consists of four identical Gaussian blobs. Each blob ends up with irregular shape and the strength of signal decreases steady from center to margin. They are placed in the center with equal distance, see Figure 5a, right. The truth contains $121 \times 4 = 484$, $169 \times 4 = 676$, and $225 \times 4 = 900$ voxels voxels based on the smoothness levels FWHM = 4 mm, 6 mm, and 8 mm accordingly. We add ES = 0.2 and 0.4 in this pattern, and further compare our method with TFCE because both methods provide an output that represents the local spatial support from nearby neighborhood. We calculate the voxel-wise TPR and TDR in this pattern.

4.1.3 | Pattern III

In particular, we want to test the performance on images with intense and focal signal (FWHM = 8 mm) as we find out the TPR goes down due to the exclusion of small true clusters when cluster extent is the only selection criteria. We set up the underlying truth to be a squared area in the center with $N_0 = 3 \times 3 = 9$ voxels, which has a very high chance not passing the cluster-extent threshold. The signal strength is set to ES = 0.4 and 0.6. We calculate the voxel-wise TPR and TDR, and cluster-wise FWER in this pattern.

The results are listed in Tables 3–5. In Pattern I, the ICM controls the FDR and FWER significantly better than the cluster-extent threshold. Sharing the same primary threshold, the sensitivities given by our ICM and cluster-extent threshold are roughly the same, though ICM has smaller failure rate on detecting the true signals when ES and sample size are small, which leads to a slightly higher TPR in some cases. In Pattern II, the ICM has higher sensitivity when the smoothness level is low (e.g., FWHM = 4 mm). At medium to



(a) Underlying truth for Pattern I (left) and II (right)

(b) An example of each smoothed image for Pattern I (left) and II (right)

FIGURE 5 Underlying truth is displayed in (a). An example of corresponding smoothed image for simulation studies is displayed in (b). Each column corresponds to one setting

high ES, the sensitivity of ICM is increasing as the sample size increases. Pattern III results are shown in Table 4. The cluster-extent method shows a high failure rate and FWER while ICM retains high sensitivity and a well-controlled FWER. In general, all methods demonstrate high sensitivity when the ES is high in both Pattern I and Pattern II. The FDR increases when the ES is higher in Pattern II because the smoothed true signal can influence more neighboring voxels. The ICM outperforms popular existing methods on controlling the voxel-wise FDR and cluster-wise FWER, and can maintain a fairly well sensitivity when the smoothness level and ES are both low.

4.2 | Examining FWER when $\beta_v = 0$ for all v's

We further generate images with zero ES (the underlying truth is null) to assess the capability of controlling FWE. The voxels in the original image follow a normal distribution $N(0, 1)$, and then we apply the Gaussian filter on each image using FWHM $=$ 4, 6, and 8 mm. The FWER is controlled at $\alpha = .05$.

From the results in Table 6, the ICM controls FWER well around 5% in most situations, while the cluster-extent threshold has a much inflated FWER than the α-level.

4.3 | Evaluating the impact from various primary thresholds

Since the performance of ICM in previous analysis can be affected by the selection of primary threshold, we further evaluate ICMs performance under various primary thresholds. In the meantime, we compare the results given by cluster-extent threshold and TFCE. The underlying truth is a combination of Pattern I through Pattern III: a cluster of $7 \times 7 = 49$ voxels and a cluster of $3 \times 3 = 9$ voxels

TABLE 3 Simulation result for Pattern I: general performance under a common distribution pattern: A squared area with $N_0 = 7 \times 7 = 49$ voxels in the center is set as underlying truth

	30 per arm		60 per arm		100 per arm	
	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent
FWHM = 4 mm						
ES = 0.4						
TPR	0.890 ± 0.089	0.890 ± 0.089	0.998 ± 0.009	0.998 ± 0.009	1	1
FDR	0.009 ± 0.026	0.020 ± 0.041	0.004 ± 0.017	0.012 ± 0.030	0.005 ± 0.021	0.017 ± 0.034
FWER	8%	18%	2%	11%	2%	17%
ES = 0.6						
TPR	0.999 ± 0.002	0.999 ± 0.002	1	1	1	1
FDR	0.004 ± 0.017	0.008 ± 0.024	0	0.001 ± 0.009	0.007 ± 0.026	0.017 ± 0.039
FWER	4%	9%	0	1%	6%	17%
ES = 0.8						
TPR	1	1	1	1	1	1
FDR	0.003 ± 0.014	0.010 ± 0.027	0.002 ± 0.011	0.007 ± 0.023	0.001 ± 0.010	0.002 ± 0.013
FWER	3%	11%	3%	8%	0	1%
FWHM = 6 mm						
ES = 0.2						
TPR	0.560 ± 0.173	0.540 ± 0.182	0.929 ± 0.101	0.928 ± 0.106	0.998 ± 0.009	0.998 ± 0.009
FDR	0.021 ± 0.066	0.022 ± 0.065	0.006 ± 0.032	0.017 ± 0.054	0.005 ± 0.022	0.011 ± 0.037
FWER	9%	10%	4%	6%	2%	6%
ES = 0.4						
TPR	0.999 ± 0.005	0.999 ± 0.005	1	1	1	1
FDR	0.001 ± 0.004	0.005 ± 0.026	0.005 ± 0.023	0.006 ± 0.027	0.008 ± 0.033	0.011 ± 0.039
FWER	0	3%	2%	3%	4%	6%
ES = 0.6						
TPR	1	1	1	1	1	1
FDR	0	0.006 ± 0.030	0.004 ± 0.023	0.012 ± 0.043	0.002 ± 0.014	0.008 ± 0.035
FWER	0	4%	1%	6%	1%	5%
FWHM = 8 mm						
ES = 0.2						
TPR	0.847 ± 0.187	0.856 ± 0.170	0.999 ± 0.008	0.999 ± 0.008	1	1
FDR	0.006 ± 0.038	0.006 ± 0.040	0.005 ± 0.028	0.017 ± 0.060	0.007 ± 0.043	0.015 ± 0.061
FWER	2%	2%	2%	7%	2%	6%
ES = 0.4						
TPR	1	1	1	1	1	1
FDR	0.013 ± 0.055	0.029 ± 0.075	0.006 ± 0.033	0.018 ± 0.057	0.008 ± 0.042	0.012 ± 0.052
FWER	5%	13%	2%	8%	2%	4%
ES = 0.6						
TPR	1	1	1	1	1	1
FDR	0.008 ± 0.043	0.021 ± 0.065	0.009 ± 0.043	0.026 ± 0.073	0.007 ± 0.035	0.009 ± 0.043
FWER	3%	9%	4%	11%	2%	3%

with Pattern I distribution, a cluster of $9 \times 9 = 81$ voxels, and a cluster of $3 \times 3 = 9$ voxels with Pattern II distribution. The ES is set to 0.4, and we present the result of smoothness level at FWHM = 6 mm.

The FDR, TPR, and FWER for ICM and cluster-extent threshold at different primary threshold values are displayed in Figure 6. In general, the ICM controls the cluster-wise FWER better than cluster-extent threshold. At voxel level, the ICM has a higher sensitivity and a

TABLE 4 Simulation result for Pattern II: general performance under a common distribution pattern: Four identical Gaussian blobs with 121 × 4 = 484 (FWHM = 4 mm), 169 × 4 = 676 (FWHM = 6 mm), and 225 × 4 = 900 (FWHM = 8 mm) voxels are set as underlying truth

	30 per arm		60 per arm		100 per arm	
	p_c^{Chrf}	TFCE	p_c^{Chrf}	TFCE	p_c^{Chrf}	TFCE
FWHM = 4 mm						
ES = 0.2						
TPR	0.239 ± 0.056	0.477 ± 0.060	0.603 ± 0.040	0.849 ± 0.027	0.851 ± 0.025	0.966 ± 0.012
FDR	0.030 ± 0.030	0.030 ± 0.015	0.027 ± 0.013	0.039 ± 0.016	0.029 ± 0.014	0.054 ± 0.016
FWER	2%	4%	0	2% (2 miss out)	4%	8%
ES = 0.4						
TPR	0.709 ± 0.041	0.975 ± 0.010	0.960 ± 0.013	0.999 ± 0.002	0.996 ± 0.003	1
FDR	0.009 ± 0.006	0.063 ± 0.019	0.019 ± 0.008	0.124 ± 0.019	0.043 ± 0.011	0.187 ± 0.016
FWER	2%	4%	2%	0 (3 miss out)	0	4%
ES = 0.6						
TPR	0.970 ± 0.014	0.999 ± 0.001	0.999 ± 0.001	1	1	1
FDR	0.020 ± 0.008	0.127 ± 0.024	0.066 ± 0.014	0.221 ± 0.015	0.141 ± 0.015	0.261 ± 0.008
FWER	0	4%	2%	2%	0	2%
FWHM = 6 mm						
ES = 0.2						
TPR	0.316 ± 0.079	0.819 ± 0.053	0.745 ± 0.045	0.966 ± 0.013	0.919 ± 0.020	0.995 ± 0.004
FDR	0.007 ± 0.011	0.059 ± 0.03	0.019 ± 0.011	0.107 ± 0.027	0.034 ± 0.013	0.166 ± 0.024
FWER	0	12% (4 miss out)	0	4%	0	4%
ES = 0.4						
TPR	0.938 ± 0.012	0.997 ± 0.003	0.995 ± 0.003	0.999 ± 0.001	1	1
FDR	0.040 ± 0.013	0.177 ± 0.027	0.119 ± 0.016	0.254 ± 0.022	0.191 ± 0.012	0.286 ± 0.021
FWER	0	4%	0	2%	2%	2%
ES = 0.6						
TPR	0.996 ± 0.003	0.999 ± 0.001	1	1	1	1
FDR	0.128 ± 0.017	0.259 ± 0.020	0.219 ± 0.007	0.302 ± 0.020	0.249 ± 0.007	0.336 ± 0.018
FWER	0	4%	0	0	2%	2%
FWHM = 8 mm						
ES = 0.2						
TPR	0.626 ± 0.058	0.954 ± 0.021	0.903 ± 0.021	0.996 ± 0.003	0.974 ± 0.007	0.999 ± 0.001
FDR	0.026 ± 0.014	0.145 ± 0.037	0.048 ± 0.016	0.223 ± 0.036	0.099 ± 0.018	0.286 ± 0.028
FWER	0	2%	2%	8%	4%	12%
ES = 0.4						
TPR	0.982 ± 0.007	0.999 ± 0.001	0.999 ± 0.001	1	1	1
FDR	0.114 ± 0.017	0.297 ± 0.028	0.210 ± 0.016	0.357 ± 0.018	0.189 ± 0.013	0.284 ± 0.018
FWER	0%	4%	4%	8%	2%	4%
ES = 0.6						
TPR	0.999 ± 0.001	0.999 ± 0.001	1	1	1	1
FDR	0.220 ± 0.015	0.362 ± 0.015	0.296 ± 0.012	0.396 ± 0.010	0.346 ± 0.009	0.417 ± 0.011
FWER	5%	10%	0	2%	2%	4%

Note: Miss out counts for the number of samples that fail to detect any cluster.

lower FDR. For TFCE, the sensitivity is 0.966, the FDR equals 0.122, and the FWER is 4%. Our ICM has a consistently well control on FDR and FWER, and improved sensitivity when the primary threshold is slightly loosen.

5 | DISCUSSION

Cluster-wise inference is one of the most common approaches in fMRI analysis in recent years. We report on the development of a

TABLE 5 Simulation result for Pattern III: A squared area with $N_0 = 3 \times 3 = 9$ voxels in the center is set as underlying truth. The 95% CI is provided for FWER because in most settings under this pattern, the Cluster-extent threshold fail to detect any significant regions in some datasets

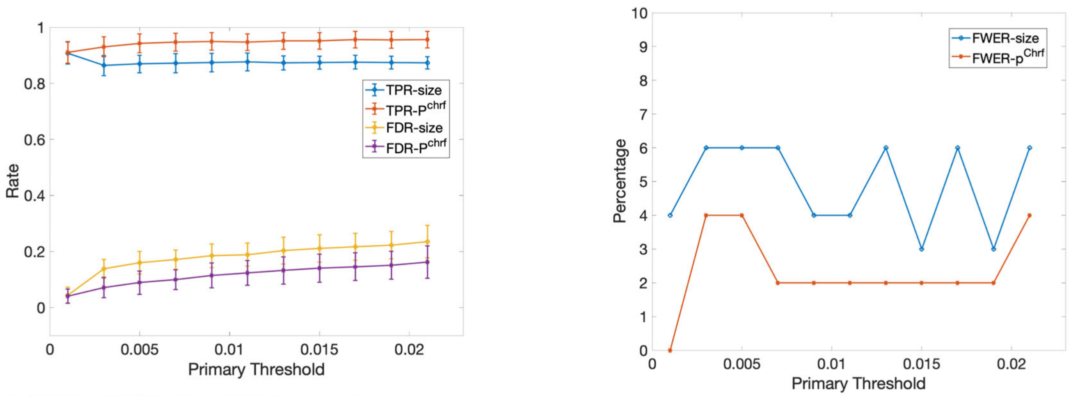
	30 per arm		60 per arm		100 per arm	
	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent
FWHM = *8 mm*						
ES = 0.4						
TPR	1	0	1	0	1	0.01 ± 0.1
FDR	0	NA[a]	0.025 ± 0.125	NA[a]	0.033 ± 0.14	0.951 ± 0.131
FWER	0	NA[a]	4%(0%, 8%)	NA[a]	5%(1%, 9%)	100%(1)[b]
ES = 0.6						
TPR	0.720 ± 0.040	0.976 ± 0.011	0.960 ± 0.012	0.801 ± 0.048	0.996 ± 0.003	0.797 ± 0.045
FDR	0.008 ± 0.005	0.062 ± 0.018	0.017 ± 0.009	0.023 ± 0.019	0.042 ± 0.011	0.036 ± 0.016
FWER	6% (1%, 11%)	91% (59%, 100%)	5% (1%, 9%)	80% (15%, 95%)	7% (2%, 12%)	100% (12)[b]

[a] No significant clusters were detected in any samples.

[b] Numbers in the parenthesis are the number of datasets that cluster-extent threshold can detect any significant clusters.

TABLE 6 Simulation result for $\beta_v = 0$

	FWHM = 4 mm		FWHM = 6 mm		FWHM = 8 mm	
	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent	p_c^{Chrf}	Cluster extent
30 per arm	0	9% (3%, 15%)	3% (0, 6%)	4% (0, 8%)	5% (1%, 9%)	6% (1%, 11%)
60 per arm	1% (0, 3%)	14% (7%, 21%)	5% (1%, 9%)	9% (3%, 15%)	1% (0, 3%)	8% (3%, 13%)
100 per arm	5% (1%, 9%)	16% (9%, 23%)	7% (2%, 12%)	11% (5%, 17%)	3% (0, 6%)	5% (1%, 9%)



(a) FDR, TPR for ICM and cluster-extent threshold at different primary threshold.

(b) The FWER for ICM and cluster-extent threshold at different primary threshold.

FIGURE 6 The FDR, TPR, and FWER for ICM and cluster-extent threshold at different primary threshold. The primary threshold ranges from 0.001 to 0.021

cluster-wise statistic to better characterize the overall statistical properties of a cluster by integrating the cluster extent, voxel-level significance, and dependence structure. Our simulation and data example show that the proposed method is computationally efficient (MATLAB execution time: ICM = 0.0228 s, cluster-extent method = 0.008 s, on 2.6 GHz 6-Core Intel Core i7) and can improve the accuracy for cluster-wise inference.

Our method makes at least two innovative contributions. First, we account for the spatial dependence of voxels with a parametric approach and incorporate this dependence structure into the Fisher's combined statistic. This procedure leads to a more accurate estimation of the combined statistic distribution. Furthermore, the Chernoff bound effectively solves the floating point problem for extremely small values of the combined statistic significance level. In sum, inference about the clusters have improved accuracy and interpretability. The simulation results show that the ICM has a higher sensitivity with well-controlled FWER among existing methods, and is applicable to a wider range of signals.

In our example involving task-based data, we identified the brain regions activated by positive RPEs in the context of an RL study. The additional activation in inferior temporal gyrus was observed under the threshold given by ICM in both full sample (52 participants) and subsample (36 participants), whereas the cluster-extent threshold or other combining *p* methods failed to detect this region in either scenario. As activation of this region is frequently reported in the context of other task-based studies in similar settings, it is possibly a biologically meaningful finding. The rs-fMRI data example with noninformative covariates provides evidence supportive of a more rigorous control by ICM on the false-positive clusters emerging from the cluster-extent threshold, which yields to a 70% decrease in the FWER compared to classical cluster-wise inference methods.

Sharing limitations with cluster-extent based method, our statistic is largely affected by the selection of the primary threshold. When an overly-liberal primary threshold is given, our statistic may generate no significant findings or fewer significant findings. This is due to the strict control of FWER. If the primary threshold is too stringent, there will be fewer significant findings, and the total "information" contains in a cluster will be similar to that of a random noise cluster. There are various ways to deal with the primary thresholding problems. An optimal primary threshold is given by eBass (Ge et al., 2021) that avoids the oversized or undersized clusters for cluster-level inference can effectively control the false discoveries and family-wise errors. The ICM can also incorporate with the TFCE (Smith & Nichols, 2009) framework, serving as a more informative replacement of cluster extent e(h) in their formulation. In summary, as the ICM is expressed in a closed form, it is compatible and can be implemented with a wide range of existing software platforms. We provide the implementation of ICM in GitHub at https://github.com/yierge/ICM.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Yunjiang Ge https://orcid.org/0000-0003-4064-8584
Gang Chen https://orcid.org/0000-0002-2960-089X
Shuo Chen https://orcid.org/0000-0002-7990-4947

REFERENCES

Alves, G., & Yu, Y.-K. (2014). Accuracy evaluation of the unified p-value from combining correlated p-values. *PLoS One*, *9*(3), e91225.

Boehme, R., Deserno, L., Gleich, T., Katthagen, T., Pankow, A., Behr, J., … Schlagenhauf, F. (2015). Aberrant salience is related to reduced reinforcement learning signals and elevated dopamine synthesis capacity in healthy adults. *Journal of Neuroscience*, *35*(28), 10103–10111.

Bowman, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *Journal of the American Statistical Association*, *102*(478), 442–453.

Bowman, F. D., Guo, Y., & Derado, G. (2007). Statistical approaches to functional neuroimaging data. *Neuroimaging Clinics of North America*, *17*(4), 441–458.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, *31*(4), 987–992.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition ii: An empirical review of 275 pet and fmri studies. *Journal of Cognitive Neuroscience*, *12*(1), 1–47.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919.

Chen, Z., Yang, W., Liu, Q., Yang, J. Y., Li, J., & Yang, M. Q. (2014). A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics*, *15*(17), 1–7.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Cicero, D. C., Kerns, J. G., & McCarthy, D. M. (2010). The aberrant salience inventory: A new measure of psychosis proneness. *Psychological Assessment*, *22*(3), 688–701.

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017). Fmri clustering in afni: False-positive rates redux. *Brain Connectivity*, *7*(3), 152–171.

Derado, G., Bowman, F. D., & Kilts, C. D. (2010). Modeling the spatial and temporal dependence in fmri data. *Biometrics*, *66*(3), 949–957.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, *102*(477), 93–103.

Eklund, A., Knutsson, H., & Nichols, T. E. (2019). Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Human Brain Mapping*, *40*(7), 2017–2032.

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*, *113*(28), 7900–7905.

Fisher, R. A. (1992). Statistical methods for research workers. *Breakthroughs in statistics* (66–70). New York, NY: Springer.

Ge, Y., Hare, S., Chen, G., Waltz, J. A., Kochunov, P., Elliot Hong, L., & Chen, S. (2021). Bayes estimate of primary threshold in clusterwise functional magnetic resonance imaging inferences. *Statistics in Medicine*, *40*(25), 5673–5689.

Gopinath, K., Krishnamurthy, V., Lacey, S., & Sathian, K. (2018). Accounting for non-gaussian sources of spatial correlation in parametric functional magnetic resonance imaging paradigms II: A method to obtain first-level analysis residuals with uniform and gaussian spatial autocorrelation function and independent and identically distributed time-series. *Brain Connectivity*, *8*(1), 10–21.

Hare, S. M., Adhikari, B. M., Du, X., Garcia, L., Bruce, H., Kochunov, P., … Hong, L. E. (2021). Local versus long-range connectivity patterns of auditory disturbance in schizophrenia. *Schizophrenia Research*, *228*, 262–270.

Hare, S. M., Ford, J. M., Ahmadi, A., Damaraju, E., Belger, A., Bustillo, J., … Functional Imaging Biomedical Informatics Research Network. (2017). Modality-dependent impact of hallucinations on low-frequency fluctuations in schizophrenia. *Schizophrenia Bulletin*, *43*(2), 389–396.

Hayasaka, S., & Nichols, T. E. (2004). Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, *23*(1), 54–63.

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, *105*(1), 239–246.

Herath, P., Kinomura, S., & Roland, P. E. (2001). Visual recognition: Evidence for two distinctive mechanisms from a pet study. *Human Brain Mapping*, *12*(2), 110–119.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379–9384.

Kochunov, P., Lancaster, J. L., Thompson, P., Woods, R., Mazziotta, J., Hardies, J., & Fox, P. (2001). Regional spatial normalization: Toward an optimal target. *Journal of Computer Assisted Tomography*, *25*(5), 805–816.

Kost, J. T., & McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, *60*(2), 183–190.

Krylov, V. I., & Stroud, A. H. (2006). *Approximate calculation of integrals*. Chelmsford, MA: Courier Corporation.

Lancaster, J. L., Cykowski, M. D., McKay, D. R., Kochunov, P. V., Fox, P. T., Rogers, W., … Mazziotta, J. (2010). Anatomical global spatial normalization. *Neuroinformatics*, *8*(3), 171–182.

Lancaster, J. L., Laird, A. R., Eickhoff, S. B., Martinez, M. J., Fox, P. M., & Fox, P. T. (2012). Automated regional behavioral analysis for human brain images. *Frontiers in Neuroinformatics*, *6*, 23.

Lazar, N. A., Luna, B., Sweeney, J. A., & Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, *16*(2), 538–550.

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, *105*(48), 18718–18723.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, *23*(4), 439–464.

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

Liu, Y., & Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, *115*(529), 393–402.

Mesulam, M.-M. (1998). From sensation to cognition. *Brain: A Journal of Neurology*, *121*(6), 1013–1052.

Minasny B., & McBratney A. B. (2005). The Matérn function as a general model for soil variograms. *Geoderma*, *128*(3-4), 192–207. https://doi.org/10.1016/j.geoderma.2005.04.003

Murray, G., Corlett, P., Clark, L., Pessiglione, M., Blackwell, A., Honey, G., … Fletcher, P. (2008). Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Molecular Psychiatry*, *13*(3), 267–276.

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446.

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

Onitsuka, T., Shenton, M. E., Salisbury, D. F., Dickey, C. C., Kasai, K., Toner, S. K., … McCarley, R. W. (2004). Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: An MRI study. *American Journal of Psychiatry*, *161*(9), 1603–1611.

Park, B.-Y., Byeon, K., & Park, H. (2019). Funp (fusion of neuroimaging preprocessing) pipelines: A fully automated preprocessing software for functional magnetic resonance imaging. *Frontiers in Neuroinformatics*, *13*, 5.

Poline, J.-B., & Mazoyer, B. M. (1993). Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of Cerebral Blood Flow & Metabolism*, *13*(3), 425–437.

Poline, J.-B., Worsley, K. J., Evans, A. C., & Friston, K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, *5*(2), 83–96.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Raballo, A., Cicero, D. C., Kerns, J. G., Sanna, S., Pintus, M., Agartz, I., … Preti, A. (2019). Tracking salience in young people: A psychometric field test of the aberrant salience inventory (asi). *Early Intervention in Psychiatry*, *13*(1), 64–72.

Roiser, J. P., Howes, O. D., Chaddock, C. A., Joyce, E. M., & McGuire, P. (2013). Neural and behavioral correlates of aberrant salience in individuals at risk for psychosis. *Schizophrenia Bulletin*, *39*(6), 1328–1336.

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.

Spence, J. S., Carmack, P. S., Gunst, R. F., Schucany, W. R., Woodward, W. A., & Haley, R. W. (2007). Accounting for spatial dependence in the analysis of spect brain imaging data. *Journal of the American Statistical Association*, *102*(478), 464–473.

Stouffer, S. (1949). A study of attitudes. *Scientific American*, *180*(5), 11–15.

Takemura, H., Samejima, K., Vogels, R., Sakagami, M., & Okuda, J. (2011). Stimulus-dependent adjustment of reward prediction error in the midbrain. *PLoS One*, *6*(12), e28337.

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

Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, *35*(10), 1319–1327.

Wald, M. J., Vasilic, B., Saha, P. K., & Wehrli, F. W. (2006). Performance comparison of the spatial autocorrelation function and the mean intercept-length in the determination of trabecular bone anisotropy in the in vivo environment. *Proceedings of the International Society for Magnetic Resonance in Medicine*, *14*, 1284.

Waltz, J. A., Xu, Z., Brown, E. C., Ruiz, R. R., Frank, M. J., & Gold, J. M. (2018). Motivational deficits in schizophrenia are associated with reduced differentiation between gain and loss-avoidance feedback in the striatum. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 239–247.

Westberg, M. (1985). Combining independent statistical tests. *The Statistician*, *34*(3), 287. https://doi.org/10.2307/2987655

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(4), 1195–1200.

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. (2014). Cluster-extent based thresholding in fmri analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419.

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918.

Yan, C.-G., Wang, X.-D., Zuo, X.-N., & Zang, Y.-F. (2016). Dpabi: Data processing & analysis for (resting-state) brain imaging. *Neuroinformatics*, *14*(3), 339–351.

Yang, Z.-H., & Chu, Y.-M. (2017). On approximating the modified bessel function of the second kind. *Journal of Inequalities and Applications*, *2017*(1), 1–8.

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S., & Wolfinger, R. D. (2007). Combining p-values in large-scale genomics experiments. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, *6*(3), 217–226.

Zhang, H., Nichols, T. E., & Johnson, T. D. (2009). Cluster mass inference via random field theory. *NeuroImage*, *44*(1), 51–61.

Zhu, Z., & Wu, Y. (2010). Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data. *Journal of Computational and Graphical Statistics*, *19*(1), 74–95.

How to cite this article: Ge, Y., Chen, G., Waltz, J. A., Hong, L. E., Kochunov, P., & Chen, S. (2022). An integrated cluster-wise significance measure for fMRI analysis. *Human Brain Mapping*, *43*(8), 2444–2459. https://doi.org/10.1002/hbm.25795

APPENDIX

(A) Parametric covariance matrix for r_{jk}

The numerical approximation of $a\chi_f^2$ requires the correlations across all pairs of the voxels within the cluster c. Commonly, a correlation matrix can be estimated empirically. However, when the matrix is large, the empirical calculation may become strenuous and inaccurate. In addition, the Pearson's correlation may not be suitable to describe the nonlinear correlation decay with distance in the smoothed fMRI images (Bowman, 2007; Minasny & McBratney, 2005; Spence et al., 2007). Here, we provide a parametric approach in estimating the spatial correlations based on the $\mathbf{\Sigma_c}$.

In Equation (1), it contains $r_{jk} = corr(x_j, x_k)$ to represent the correlation between j-th and k-th voxels in a cluster. To measure the spatial variation by the relative location of voxels, we use the Matérn covariance to calculate each specific correlation between a pair of voxels since it only depends on distances between points, though there are other parametric spatial correlation model available to describe the functional similarity between voxels (Bowman, 2007). The original form of the stationary, isotropic (we consider the Euclidean distance between voxels) Matérn covariance function is given by

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{d}{\rho}\right) \tag{A1}$$

where ρ is the spatial range parameter, d is the Euclidean distance between two voxels, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. Specifically, $K_\nu(x) = \frac{\pi(I_{-\nu}(x) - I_\nu(x))}{2\sin(\nu\pi)}$, and $I_\nu(x) = \sum_{n=0}^{\infty} \frac{1}{n!\Gamma(\nu+n+1)}\left(\frac{x}{2}\right)^{2n+\nu}$ is a particular solutions of the second-order differential equation $x^2y''(x) + xy'(x) - (x^2 + \nu^2)y(x) = 0$ (Yang & Chu, 2017).

This covariance function has great flexibility in practice because when $\nu = 1/2 + p$ for $p \in \mathcal{N}^+$, $C_\nu(d)$ can be simplified to the product of an exponential and a polynomial of order p. With this property, one can adjust the parameters for different smoothness levels. In particular, we choose the $\nu = 3/2$ so that the calculation is less complex while the function is still one time differentiable.

Let d_{jk} denotes the Euclidean distance between voxel j and voxel k, then the spatial correlation r_{jk} derived from Equation (A1) can be written into a product of an exponential function and a polynomial of order one:

$$r_{jk} = \left(1 + \frac{\sqrt{3}d_{jk}}{\rho}\right)\exp\left(-\frac{\sqrt{3}d_{jk}}{\rho}\right) \tag{A2}$$

where ρ is the characteristic length scale that measures the relevance between two voxels. It can be theoretically calculated through the spatial autocorrelation (Park, Byeon, & Park, 2019; Wald, Vasilic, Saha, & Wehrli, 2006; Zhu & Wu, 2010), or be empirically obtained by simulations. The recommended empirical values for ρ based on the full width at half maximum (FWHM) are listed in Table A1. In fMRI data analyses, $\rho = 6$ or 8 are the most commonly used parameters. We also provide a function in the GitHub package for users to customize this parameter.

(B) Proofs

B.1 Proof for Lemma 1

Proof. For cluster c, the Chernoff bound for the p-value of its combined statistic is given by $p_c^{Chrf} = \left(\frac{T_c}{f}\right)^{\frac{f}{2}} \exp\left[\frac{f}{2}\left(1 - \frac{T_c}{f}\right)\right]$.

Let $t_c = \frac{T_c}{f}$. Since f is fixed, for p_c^{Chrf} we only need to consider the function

$$h(t) = t\exp(1-t) \tag{A3}$$

The function $h(t)$ is monotone increasing for $t \in (0, 1)$ and monotone decreasing for $t \in (1, \infty)$.

Since the statistic T_c is defined by $T_c = \frac{\Psi_c}{a}$, then $\frac{T_c}{f} = \frac{\Psi_c}{af} = \frac{\Psi_c}{2n_c}$, which is the ratio of sum statistic Ψ_c and cluster size n_c. The uncorrected primary threshold is 0.05, thus we assume the primary threshold is no greater than 0.05. Then for all p_j's, $-\ logp_j \geq 3$, $\frac{min\{-\log p_j\}}{1} \geq 3$ which grantees $t > 1$.

Thus, the domain of $h(t)$ is $[3, \infty)$. $h(t)$ is always a monotone decreasing function on its domain.

Suppose the two clusters c_1, c_2 contain n_1, n_2 voxles, with sum statistic Ψ_{c_1}, Ψ_{c_2}, accordingly. Their significant level of combined statistics are $p_{c_1}^*, p_{c_2}^*$. The dfs satisfy $f_1 = f_2 = f$, where f is a constant.

If $p_{c_1}^* < p_{c_2}^*$, then by Equation (A3), $T_{c_1} > T_{c_2}$. Thus, for $t_1 = \frac{T_1}{f}, t_2 = \frac{T_2}{f}$, we have $t_1 > t_2$. Since $h(t)$ is a monotone decreasing function, the Chernoff bound for the p values of combined statistic satisfy

$$\left(\frac{T_1}{f}\right)^{\frac{f}{2}} \exp\left[\frac{f}{2}\left(1 - \frac{T_1}{f}\right)\right] < \left(\frac{T_2}{f}\right)^{\frac{f}{2}} \exp\left[\frac{f}{2}\left(1 - \frac{T_2}{f}\right)\right] \tag{A4}$$

TABLE A1 ρ for calculating the pairwise correlation r_{jk}

	Effect size				
	0.2	**0.4**	**0.6**	**0.8**	**1.0**
FWHM					
4 mm	–	2	2.5	3	3.5
6 mm	3	4	4.5	5.5	6
8 mm	4	6	7	8	9

which is $p_{c_1}^{Chrf} < p_{c_2}^{Chrf}$.

B.2 Proof for Theorem 1

Proof. When the *df*s are fixed, p_c^{Chrf} can be written in a function form as Equation (A3), where $t \in [3, \infty)$. For two clusters c_1, c_2 with s_{c_1}, s_{c_2} correspondingly. Let $s_{c_1} < s_{c_2}$, which leads to $t_{c_1} < t_{c_2}$. Thus

$$h(t_{c_1}) > h(t_{c_2}) \Rightarrow p_{c_1}^{Chrf} > p_{c_2}^{Chrf}.$$

B.3 Proof for Theorem 2

Proof. For a cluster c, since the ratio $s_c = \frac{\Psi_c}{n_c}$ is fixed, we treat it as a constant *s*, $s \geq 3$. Then the p_c^{Chrf} can be written in the same form of Equation (A3) to the power of $\frac{f}{2}$. That is $h(t)^{\frac{f}{2}}$, where $t = \frac{r}{2}$. The codomain of $h(t)$ is (0, 1).

For two clusters c_1 and c_2 with fixed *s*, let $y = h(t)$, *y* is a constant in (0, 1). If $f_1 < f_2$, we have

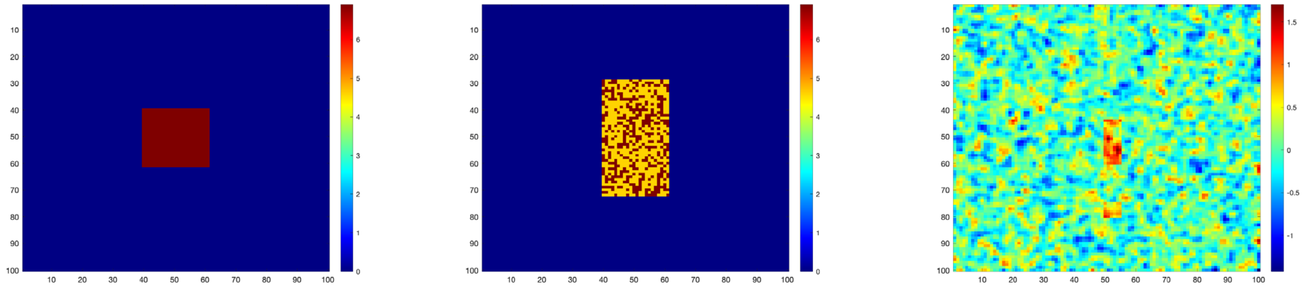
$$y^{f_1} > y^{f_2} \Rightarrow p_{c_1}^{Chrf} > p_{c_2}^{Chrf}.$$

B.4 Proof for Proposition 1

Proof. For two clusters c_1, c_2 with same ratio *s* and cluster size *n*, their approximated *df*s are given by $f_{c_1} = \frac{8n^2}{4n+\rho_{c_1}}, f_{c_2} = \frac{8n^2}{4n+\rho_{c_2}}$. Suppose $\rho_{c_1} < \rho_{c_2}$, then we have $f_{c_1} > f_{c_2}$. By the property of Equation (A3), $h(t)^{\frac{f_{c_1}}{2}} < h(t)^{\frac{f_{c_2}}{2}} \Rightarrow p_{c_1}^{Chrf} < p_{c_2}^{Chrf}$.

(C) A simulated example for Theorem 1 conclusion

Consider a 100 × 100 2D image contains a cluster consists of 22 × 22 = 484 'true voxels' with strong signal (Figure A1a). Another image of same size contains the cluster consists of the original 484 voxels plus another 484 supra-threshold but noise voxels with larger *p* values (Figure A2b). In cluster extent permutation test, the Figure A1a cluster has lower probability survive the cluster-extent threshold than Figure A1b cluster. For p_c^{Chrf}, the average signal strength of cluster s_c in Figure A1a is larger than that in Figure A1b. When their approximated *df* are about equivalent, by Proposition 1, the Figure A1b will yield to a larger p_c^{Chrf} (less significant). To show the average signal strength for approximated same *df*, we generate the third image Figure A1c that contains two clusters with same *df* and cluster size similar to the clusters in Figure A1a,b, respectively. This reflect an important feather that if the true activated region is naturally small, the p_c^{Chrf} can move up its rank among all clusters. For example, the activation region toward a stimuli is on amygdala. The cluster size will be small and has lower chance to survive the cluster-extent threshold. With moderate signal strength, such a activated region will have much higher chance to survive the ICM threshold.



(a) Strong average signal (b) Extended cluster extent (c) Data example: equal f

FIGURE A1 Images showing the relationship between average signal strength and p_c^{Chrf}. (a) contains 22 × 22 = 484 voxels with $p = .001$ (b) contains 44 × 22 = 968 voxels. 484 with $p = .001$ and 484 with $p = .01$. (c) is an example of signal strength for different cluster sizes in achieving the same *f*. The large cluster contains 102 voxels, while the small cluster contains 36 voxles