# Valid and powerful group statistics for decoding accuracy: Information Prevalence Inference using the i-th order statistic (*i*-test)

Satoshi Hirose[a,b]

[a]Center for Information and Neural Networks, National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita, Osaka 565-0871, Japan

[b]Graduate School of Medicine & Graduate School of Frontier Biosciences, Osaka University, 2-15 Yamadaoka, Suita, Osaka 565-0871, Japan

Corresponding author:
Satoshi Hirose
Center for Information and Neural Networks
National Institute of Information and Communications Technology
1-4 Yamadaoka, Suita, Osaka 565-0871, Japan
Tel: +81-80-9045-9529
Fax: +81-6-7174-8612
E-mail: satoshi.hirose@nict.go.jp

Declarations of interest:
None.

Abbreviations:

D-Acc: decoding accuracy

**Abstract**

In fMRI decoding studies using pattern classification, a “second-level” group statistical test is typically performed after “first-level” decoding analyses for individual participants. Neuroscientists often test the mean decoding accuracy across participants against the chance-level accuracy (e.g., one-sample Student *t*-test) to verify whether brain activation includes information about the label (i.e., cognitive content). However, Allefeld et al., (2016) notified that positive results for such tests only indicate that "there are some people in the population whose fMRI data carry information about the experimental condition." Thus, such tests cannot provide inferences about the trend of the majority. They proposed an alternative method, in which prevalence inference is implemented. In this study, I extended their method and propose statistical test “information prevalence inference using the i-th order statistic (*i*-test)”. Compared with the method proposed in the previous study, *i*-test has high statistical power to provide an inference regarding major population trends. In *i*-test, the i-th lowest sample decoding accuracy (the i-th order statistic) is compared to the null distribution to test whether the proportion of the higher-than-chance decoding accuracy in the population (information prevalence) is higher than the threshold. Thus, a significant result of the *i*-test can be interpreted as “*the majority of the population has information about the label in the brain*.” Numerical calculation identified its high statistical power. Also, theoretical detail is provided, and the use of this method in an fMRI decoding study is demonstrated.

Allefeld, C., Görgen, K., Haynes, J.D., 2016. Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. Neuroimage.

**Keywords**

decoding accuracy, group statistical test, information-like measure, order statistic, information prevalence inference, *i*-test

## 1. Introduction

An increasing number of fMRI studies use multi-voxel pattern classification analysis, where the cognitive content (= label), such as visual input categories (Spiridon and Kanwisher 2002, Haxby et al., 2014, Nishida and Nishimoto 2018), types of movement (Gallivan et al., 2011, Nambu et al., 2015), and the effector of movement (Hirose et al., 2015, Hirose et al., 2018) is predicted (decoded) from fMRI signal patterns. The prediction's accuracy (decoding accuracy; D-Acc) is regarded as an index of the label information in the brain. In fMRI decoding studies, the D-Acc obtained in the experiment is often compared to the chance-level accuracy, that is obtained when the brain contains no information about the label (e.g., 50% for binary classifications). When the D-Acc is higher than the chance level, the brain, or a particular brain region, contains information regarding cognitive content. In real experiments, the D-Acc for multiple participants is combined to make an inference with respect to the D-Acc's population distribution.

This study concerns a statistical test for inferring the D-Acc's population from the experimental results of multiple participants. In analogy with conventional univariate analysis for voxel activity (Friston et al., 1995), the mean D-Acc across participants has been often compared to the chance level by using Student's *t*-test (e.g. Gilbert and Fung, 2018, Gallivan et al., 2011) or permutation test (Gallivan et al., 2013, Nambu et al., 2015, Stelzer et al., 2013). The principle of these analyses is that the D-Acc sample mean is a good estimator of its population mean, and if the population mean is higher than chance, the brain activation is expected to typically contain cognitive content information. However, this is a misleading interpretation. Indeed, in theory, the expectation of each participant's D-Acc has to be equal to or higher than the random classifier and cannot be lower than the chance level (Allefeld et al., 2016). Thus, very small proportion of the population with a higher-than-chance D-Acc is sufficient to make a population mean higher than chance. Therefore, the significant results of these tests cannot provide inferences regarding the population trend; they can only infer that "*there are some people in the population whose fMRI data carry information about the experimental condition*" (Allefeld et al., 2016), indicating that such tests fail to infer the population majority trend.

The above-mentioned problem is inevitable for statistical tests of the population mean. One solution is to assess the population information prevalence (i.e., the proportion of the population with a higher D-Acc than chance level). Accordingly, Allefeld et al. proposed a statistical test on the D-Acc, where the population information prevalence is targeted instead of the population mean. This was inspired by other methods evaluating the proportion of the population for second-level (multi-subject) statistical test in fMRI data analyses, such as the dynamic causal model (Stephan et al., 2009), the second-level random effect of univariate analysis (Rosenblatt et al., 2014), and conjunction analysis (Friston et al., 1999). In this method, the null hypothesis is that the "*proportion of the population that has a higher-than-chance D-Acc is not larger than the predetermined prevalence threshold (e.g., 0.5)*." Thus, if the null hypothesis is rejected, it can be concluded that "*proportion larger than the prevalence threshold (e.g., more than half) of the population has label information in the brain activity*".

The method proposed in Allefeld et al. (2016) theoretically solved the above concern and achieved a meaningful inference regarding the population trend. However, the method has a drawback. Because the lowest D-Acc among the participants is used as the representative value, the method can often fail to capture the correct population characteristics, which leads poor statistical power. This becomes serious particularly for experiments with many participants, because large number of participants increases probability of appearance of a lower outlier D-Acc, which rarely appears for each participant. This leads a counter-intuitive drawback that the statistical power decreases when the number of participants increases (see Section 3.1).

To address this issue, in the present study, I propose the extension of the previous method. The idea is to generalize Allefeld's method to the $i$-th order statistic ($i$-th lowest D-Acc). Namely, the 2nd order statistic (2nd lowest D-Acc) or higher-order statistic is used as the representative value. As a result, $i$-test becomes tolerant to lower outliers, because the $i$-th order statistic is immune to the existence of $i-1$ lower outliers. See Section 2.2.3 for the relationship between the methods proposed by Allefeld et al. (2016) and that proposed in this paper. I name the proposed method as "information prevalence inference using the $i$-th order statistic" or $i$-test.

The remainder of the paper is organized as follows: In the next section, the proposed statistical test is explained (Section 2). Then, the advantage of the method is

demonstrated by providing numerical results (Section 3). Next, the application is demonstrated using a real fMRI dataset (Section 4). Finally, the results are summarized, and the key characteristics of the method are briefly discussed (Section 5).

## 2 Information prevalence inference using the *i*-th order statistic (*i*-test)

In this section, the statistical model is clarified, the problem is defined, and the underlying theory is outlined (Section 2.0). Next, the theoretical detail of *i*-test is explained (Section 2.1). Then, essential theoretical characteristics of *i*-test are introduced, as well as the relationship with the previous method (Section 2.2). Finally, the practical procedure for performing *i*-test is explained (Section 2.3). Section 2.3 is self-contained, to some extent, and contains minimum equations, so that readers can skip 2.1 and 2.2 to see the practical procedure before the theoretical detail.

### *2.0 notations, problem definition, and outline of the calculation*

Figure 1A shows the statistical model of the population ($\Omega$) and how experimental results (sample D-Acc in a real experiment) are obtained. Population $\Omega$ can be partitioned into two subgroups, $\Omega_+$ and $\Omega_-$. People in $\Omega_+$ have label information in their brain activity and, thus, expectation of D-Acc higher than chance, while people in $\Omega_-$ do not. When a person with index $n$ is randomly chosen from the population, he/she belongs to $\Omega_+$ ($n \in \Omega_+$), with a probability $\gamma$, or otherwise belongs to $\Omega_-$ ($n \in \Omega_-$). The experiment with $N$ participants is expressed as a two-step random sampling. First, $N$ participants are randomly sampled from the population independently of each other ("Random sampling of participant"). Each sampled participant is associated with the probability distribution (probability mass function) of the D-Acc ($p(a_n)$). Note that in this paper, lowercase $p$ is used for probability mass functions, while upper case $P$ is used for probabilities.

Second, the sample D-Acc ($\hat{a}_n$) is randomly sampled from the distribution for each participant (Random sampling of D-Acc). The objective of *i*-test is to test whether $\gamma$ is larger than the predetermined threshold $\gamma_0$ with the significance threshold $\alpha$ from $N$ participants' experimental results.

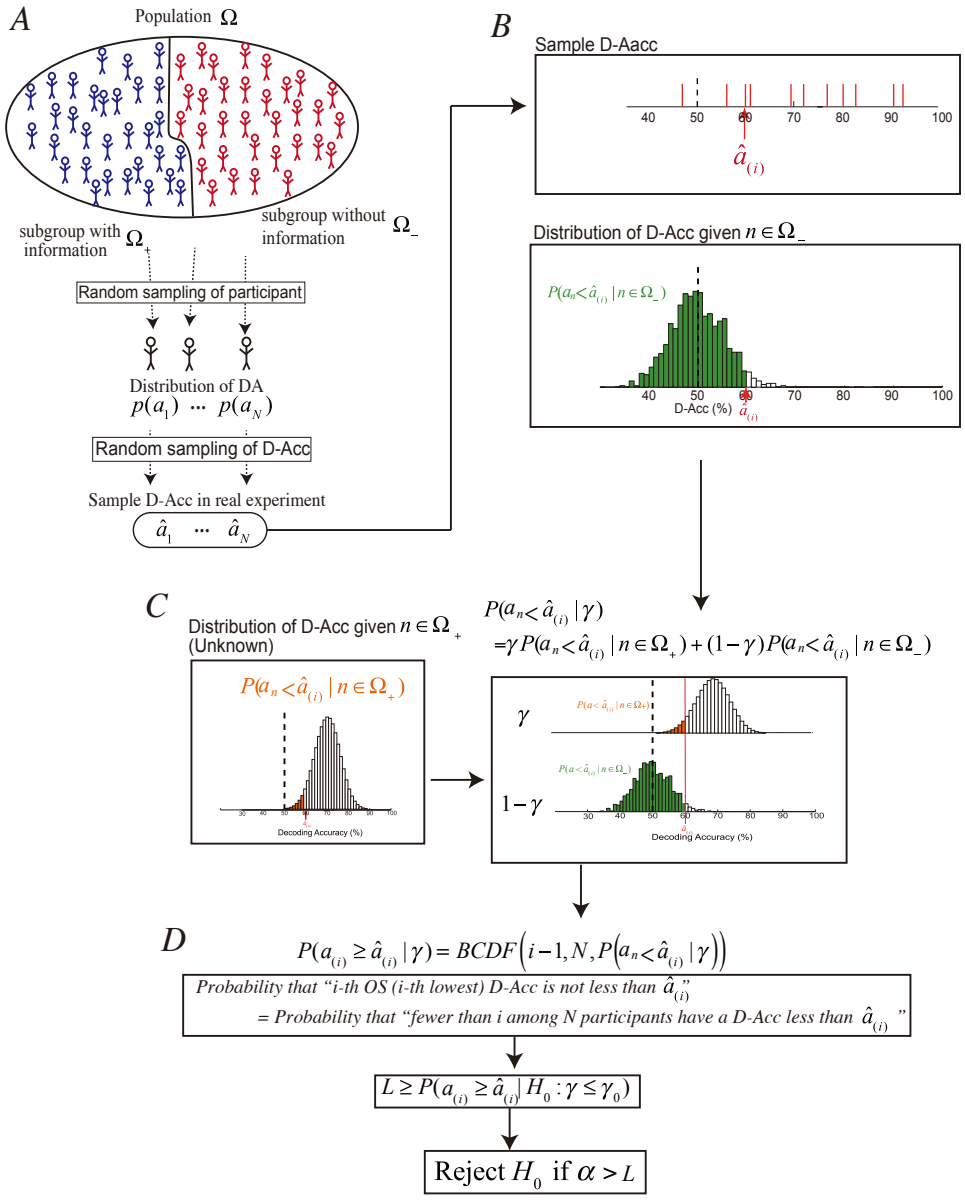



**Figure 1:** A statistical model and schematic design of $i$-test. **A:** Statistical model. The population ($\Omega$) is composed of two subgroups, $\Omega_+$, with label information, and $\Omega_-$, without label information. An experiment is expressed as two-step random sampling; first, each participant is randomly chosen from the population (Random sampling of participant), and then sample D-Acc in the experiment ($\hat{a}_n$) is randomly chosen from the distribution for each chosen participant (Random sampling of D-Acc). **B:** Values obtained from experimental results. From the experimental results, $i$-th order statistic ($i$-th lowest D-Acc; $\hat{a}_{(i)}$) is chosen (top). Also, we estimate the probability distribution given that a participant does not have label information ($p(a_n|n \in \Omega_-)$; bottom). Green

bars indicate $P(a_n < \hat{a}_{(i)} | n \in \Omega_-)$. **C:** Left: The probability distribution given that a participant has label information ($p(a_n | n \in \Omega_+)$). Orange bars indicate $P(a_n < \hat{a}_{(i)} | n \in \Omega_+)$, whose value is unknown. Right: the probability that a participant provides D-Acc less than $\hat{a}_n$ given $\gamma$, i.e. $P(a_n < \hat{a}_{(i)} | \gamma)$, can formulate as the weighted sum of $P(a_n < \hat{a}_{(i)} | n \in \Omega_+)$ and $P(a_n < \hat{a}_{(i)} | n \in \Omega_-)$ with the weight of the probability that a participant belongs to $\Omega_+$ $(\gamma)$ and to $\Omega_-$ $(1 - \gamma)$, respectively. **D**: The probability that "*$i$-th order statistic is not less than observed one* ($\hat{a}_{(i)}$)" is equivalent to the probability that "*fewer than $i$ among $N$ participants have a D-Acc less than* $\hat{a}_{(i)}$," which can be formulated with binomial cumulative distribution function. Then, we can calculate the upper bound ($L$) of the probability that "*$i$-th order statistic is not less than the observed one, given the null hypothesis*:" $P(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 : \gamma \leq \gamma_0)$. If this is smaller than the statistical threshold $\alpha$, then, we reject the null hypothesis and conclude that the population prevalence $\gamma$ is larger than $\gamma_0$.

For the statistical test, we first choose the $i$-th order statistic ($i$-th lowest D-Acc; $\hat{a}_{(i)}$) from the sample D-Acc observed in the experiment (Figure 1B top). In addition, we estimate the probability mass function of D-Acc given that the participant does not have label information ($p(a_n|n \in \Omega_-)$; Figure 1B bottom). This is achieved by empirical procedure, such as permutation test, or with assumption of distribution of D-Acc, such as binomial distribution (See Section 2.2.2). Without estimating the D-Acc probability distribution given a person has label information ($p(a_n|n \in \Omega_+)$; Figure 1C left), we formulate the probability that "*a participant provided a D-Acc less than* $\hat{a}_{(i)}$" under a given $\gamma$ ($P(a_n < \hat{a}_{(i)}|\gamma)$; Figure 1C right) as a weighted sum of $P(a_n < \hat{a}_{(i)}|n \in \Omega_+)$ and $P(a_n < \hat{a}_{(i)}|n \in \Omega_-)$ with the weight of the probability that a participant belongs to $\Omega_+$ ($\gamma$) and to $\Omega_-$ ($1-\gamma$), respectively. Then, the probability that "*$i$-th order statistic is not less than observed one* ($\hat{a}_{(i)}$)" under a given $\gamma$ is formulated ($P(a_{(i)} \geq \hat{a}_{(i)}|\gamma)$; Figure 1D top). This is equal to the probability that "*fewer than $i$ among $N$ participants have a D-Acc less than* $\hat{a}_{(i)}$." The equivalence of these two events is easily realized by finding that $i$-th lowest D-Acc is not less than $\hat{a}_{(i)}$, if and only if the number of participants with D-Acc less than $\hat{a}_{(i)}$ is fewer than $i$. Although we cannot calculate the probability directly, we can calculate its upper bound ($L$) under the null hypothesis of "*$\gamma$ is at or less than the predetermined threshold* $\gamma_0$" ($\mathrm{H}_0: \gamma \leq \gamma_0$). Then, when the upper bound is smaller than $\alpha$, we can reject the null hypothesis and prove that $\gamma$ is larger than the predetermined threshold $\gamma_0$ with the significance threshold $\alpha$.

The notations are tabulated in Table 1.

Table 1: Notations used in the present study.

| |
|---|
| Predetermined Constant |
| $N$: Number of participants |
| $\gamma_0$: prevalence threshold |
| $\alpha$: statistical threshold |
| $i$: rank of order statistic |
| |
| Unknown population value |
| $\gamma$: population information prevalence, i.e. $P(n \in \Omega_+)$ |
| |
| Population |
| $\Omega$: whole population |
| $\Omega_+$: population subgroup with label information |
| $\Omega_-$: population subgroup without label information |
| |
| Probabilistic Variable |
| $a_n$: D-Acc of a participant with index $n$ |
| $a_{(i)}$: $i$-th order statistic of D-Acc from $N$ participants |
| |
| Experimental results |
| $\hat{a}_n$: observed D-Acc of participants in the experiment ($n = 1 \dots N$) |
| $\hat{a}_{(i)}$: the i-th order statistic of observed D-Acc in the experiment |
| |
| Function |
| $BPDF(k, K, \rho) = {}_K C_k \rho^k (1-\rho)^{(K-k)}$ : binomial probability mass function |
| $BCDF(k, K, \rho) = \sum_{h=0}^{k} BPDF(h, K, \rho)$: binomial cumulative probability mass function |

*2.1 Theoretical details*

As noted above, the objective of *i*-test is to test whether $\gamma$ is larger than the predetermined threshold $\gamma_0$. Thus, the alternative hypothesis of $\mathrm{H}_1: \gamma > \gamma_0$ is set, and the null hypothesis is its negation $\mathrm{H}_0: \gamma \leq \gamma_0$. Under this null hypothesis, we evaluate whether the $i$-th lowest sample D-Acc $\hat{a}_{(i)}$ is so high that it rarely (with a probability less than the significance threshold $\alpha$) occurs.

This calculation is conducted in four steps. First, we formulated the probability that "*each participant's D-Acc is less than* $\hat{a}_{(i)}$" under a given $\gamma$ ($P\left(a_n < \hat{a}_{(i)} | \gamma\right)$); eq. (2.2)]. Second, the probability that "*the* $i$*-th lowest D-Acc is not less than the one actually observed*" under a given $\gamma$ ($P\left(a_{(i)} \geq \hat{a}_{(i)} | \gamma\right)$) is formulated (eq. (2.3)). Third, the lower bound ($Q_n$) of the probability that "*each participant's D-Acc is less than* $\hat{a}_{(i)}$" under the null hypothesis ($P\left(a_n < \hat{a}_{(i)} | \mathrm{H}_0: \gamma \leq \gamma_0\right)$) is calculated (eqs. (2.6), (2.7)). Fourth, the upper bound ($L$) of the probability that "*the* $i$*-th lowest D-Acc is not less than the one actually observed*" under the null hypothesis ($P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0: \gamma \leq \gamma_0\right)$) is calculated (eqs. (2.9), (2.10)). And finally, $L$ is compared to $\alpha$.

Formulation of $P(a_n < \hat{a}_{(i)} | \gamma)$

Considering that a participant belongs to $\Omega_+$ with the probability $\gamma$, $p(a_n|\gamma)$ is formulated as follows:

$$p(a_n|\gamma) = (1-\gamma)\, p(a_n | n \in \Omega_-) + \gamma\, p(a_n | n \in \Omega_+) \qquad (2.1),$$

where the D-Acc probability distribution (probability mass function) given a person has and doesn't have label information is expressed as $p(a_n | n \in \Omega_+)$ and $p(a_n | n \in \Omega_-)$, respectively. From Eq. (2.1), the probability that *"D-Acc of a participant is less than* $\hat{a}_{(i)}$*"* is formulated as follows:

$$P\left(a_n < \hat{a}_{(i)} | \gamma\right) = (1-\gamma) P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right) + \gamma\, P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_+\right) \qquad (2.2).$$

Here, we assume the identical distribution among participants, i.e. $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_+\right)$ and $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ are independent of $n$, and therefore,

$P(a_n < \hat{a}_{(i)}|\gamma)$ is independent of $n$ (see Appendix A for the relaxation of this assumption).

<u>Formulation of $P(a_{(i)} \geq \hat{a}_{(i)}|\gamma)$</u>

Then, we formulate the probability that the "*$i$-th lowest D-Acc is not less than $\hat{a}_{(i)}$*" under a given $\gamma$ as follows;

$$P(a_{(i)} \geq \hat{a}_{(i)}|\gamma) = BCDF(i-1, N, P(a_n < \hat{a}_{(i)}|\gamma)) \qquad (2.3).$$

This is because the probability that "*$i$-th lowest D-Acc is not less than $\hat{a}_{(i)}$*" is equivalent to the probability that "fewer than $i$ *among* $N$ *participants have a D-Acc less than* $\hat{a}_{(i)}$."

<u>Lower bound of $P(a_n < \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0)$</u>

By only noting the basic nature of probability, i.e. $P(a_n < \hat{a}_{(i)}|n \in \Omega_+) \geq 0$, we easily find the lower bound of $P(a_n < \hat{a}_{(i)}|\gamma)$ from eq. (2.2) as

$$P(a_n < \hat{a}_{(i)}|\gamma) \geq (1-\gamma)P(a_n < \hat{a}_{(i)}|n \in \Omega_-) \qquad (2.4).$$

Furthermore, under the null hypothesis $\mathrm{H}_0{:}\gamma \leq \gamma_0$, we can find the lower bound of the probability as

$$P(a_n < \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0) \geq (1-\gamma_0)P(a_n < \hat{a}_{(i)}|n \in \Omega_-) \qquad (2.5).$$

This is because the right side of eq. (2.4) is nondecreasing function of $\gamma$. Here, we define $Q_n$ as the right side of eq. (2.5), i.e.

$$Q_n = (1-\gamma_0)P(a_n < \hat{a}_{(i)}|n \in \Omega_-) \qquad (2.6),$$

and we can reformulate eq. (2.5) as

$$Q_n \leq P(a_n < \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0) \qquad (2.7).$$

<u>Upper bound of $P(a_{(i)} \geq \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0)$</u>

Because $BCDF(i-1, N, P(a_n < \hat{a}_{(i)}|\gamma))$ in eq. (2.3) monotonically decreases with $P(a_n < \hat{a}_{(i)}|\gamma)$, and $P(a_n < \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0)$ is bounded below with $Q_n$ [Eq. (2.7)], we find that $P(a_{(i)} \geq \hat{a}_{(i)}|\mathrm{H}_0{:}\gamma \leq \gamma_0)$ is bounded above as follows;

$$BCDF(i-1, N, Q_n) \geq P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 : \gamma \leq \gamma_0\right) \quad (2.8).$$

Here, we define $L$ as the left side of eq. (2.8), i.e.

$$L = BCDF(i-1, N, Q_n) \quad (2.9).$$

and we can reformulate eq. (2.8) as

$$L \geq P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 : \gamma \leq \gamma_0\right) \quad (2.10).$$

If we know $P\left(a_n < \hat{a}_{(i)} | n \in \Omega_-\right)$, we can calculate $Q_n$ with Eq. (2.6), and consequently, we can calculate $L$ with Eq. (2.9). When $L$ is smaller than the statistical threshold $\alpha$, it is proven that $P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 : \gamma \leq \gamma_0\right)$ is smaller than $\alpha$ from Eq. (2.10). Thus, we can conclude that the event that the "*$i$-th lowest sample D-Acc is not less than the observed one* $\hat{a}_{(i)}$" rarely (with the probability of less than $\alpha$) occurs under the null hypothesis. Therefore, we can reject the null hypothesis and accept $\mathrm{H}_1 : \gamma > \gamma_0$, i.e., "*proportion larger than* $\gamma_0$ *of the population has label information in the brain activity.*"

## *2.2 Theoretical characteristics*

### *2.2.1 Predetermined parameters*

*i*-test requires three predetermined parameters: $\alpha$ (threshold for statistical significance), $\gamma_0$ (threshold for the information prevalence), and $i$ (rank of the order statistic). Together with the number of participants $N$, they should satisfy the following inequality;

$$\alpha > BCDF(i-1, N, 1-\gamma_0) \quad (2.11).$$

This constraint is derived as follows: From Eq. (2.6) and noting that $P\left(a_n < \hat{a}_{(i)} | n \in \Omega_-\right) \leq 1$, $Q_n$ is bounded above with $1-\gamma_0$. Thus, from Eq. (2.9), and because $L$ decreases with $Q_n$, $L$ is bounded below as follows:

$$L \geq BCDF(i-1, N, 1-\gamma_0). \quad (2.12),$$

Thus, Eq. (2.11) should be satisfied; otherwise, $L$ is always at or larger than $\alpha$ and the statistical test never provides a positive result.

See Section 2.3 Step 1 for the empirical guidelines for choosing particular parameter values within this constraint.

### *2.2.2 Estimation of $P\left(a_n < \hat{a}_{(i)} | n \in \Omega_-\right)$*

For the calculation of $L$, we need to estimate the probability that "*a participant provided a D-Acc less than* $\hat{a}_{(i)}$" given that a participant does not have label information in the brain, i.e. $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$. This is achieved by estimation of probability mass function $p(a_n | n \in \Omega_-)$, by means of empirical procedures such as permutation test (Good, 2000, Ojala and Garriga 2010), or with assumption of its distribution, e.g. binomial distribution. *i*-test can be combined with any kinds of the estimation procedure and this study does not cover the validation of the estimation. But, for the empirical use, I introduce two commonly-used methods.

<u>*Permutation test*</u>

In the permutation test, the labels for the trials are randomly permuted and generate permuted dataset. Then, the decoding analysis is applied to this permuted dataset and calculate the decoding accuracy with the same procedure used in the main analysis. This permutation is repeated multiple times and the distribution of D-Acc from the permutations are regarded as the D-Acc given that there is no label information in the brain.

When $M$ permutations are performed for each participant, we calculate D-Acc with each of the permuted data. Then, we estimate $P\left(a < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ by calculating the proportion of permutations that provide performance less than $\hat{a}_{(i)}$,

$$P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right) = \frac{1}{M \times N} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[\hat{a}_{perm(n,m)} < \hat{a}_{(i)}\right] \qquad (2.13).$$

where $\hat{a}_{perm(n,m)}$ is the D-Acc obtained from the m-th permutation of the n-th participant and Iverson bracket $[\cdot]$ has the value 1 for a true condition and the value 0 for a false condition.

<u>*Assumption of the binomial distribution*</u>

If we assume that the decoder predicted the label randomly and independently in every trials, the D-Acc given the brain activation does not include label information follows the binomial distribution

$$p(a_n | n \in \Omega_-) = BPDF(a_n \times \mathrm{N}_{trial}, \mathrm{N}_{trial}, 0.5) \qquad (2.14).$$

Under this assumption, we estimate $P\left(a < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ as

$$P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right) = BCDF(a_n \times \mathrm{N}_{trial} - 1\,, \mathrm{N}_{trial}\,, 0.5) \qquad (2.15).$$

*2.2.3 Relation with the previous method in Allefeld et al., (2016)*

By fixing $i = 1$, Eq. (2.9) becomes simply the $N$-th power of $1 - Q_n$;

$$L = (1 - Q_n)^N \qquad (2.16).$$

By doing so, *i*-test comes down to the method previously proposed by Allefeld et al., (2016), referred to as the "Permutation-based information prevalence inference using the minimum statistic." Hereafter, I term this as *i*-test-one.

*2.2.4 Possible extension for relaxing the identical distribution among participants*

I assume the identical distribution of D-Acc among participants for the derivation of *i*-test. But strictly speaking, this assumption is often broken in a real fMRI situation. Therefore, I propose a possible extension of *i*-test with the relaxation of this assumption. See Appendix A for the derivation of this extension.

*2.3 Practical Procedure*

The procedure for performing the *i*-test is constructed with 1) set three predetermined parameters: $\alpha$ (statistical threshold), $\gamma_0$ (prevalence threshold) and $i$ (rank of the order statistic), 2) identify $\hat{a}_{(i)}$ and estimate the probability that "*a participant provided a D-Acc less than* $\hat{a}_{(i)}$" given that a participant does not have label information in the brain, i.e. $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ and 3) calculate the upper bound of the "*i-th lowest sample D-Acc is at or higher than observed one* $\hat{a}_{(i)}$;" $P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 : \gamma \leq \gamma_0\right)$ and compare it with $\alpha$.

Step 1: Determine the values of the three predetermined parameters

The *i*-test's positive results can test the hypothesis that "*proportion larger than* $\gamma_0$ *of the population has label information in the brain activity"* with a significant threshold of α. In the first step of the *i*-test, we determined the three parameters; $\alpha$ (the threshold for statistical significance), $\gamma_0$ (the threshold for the information prevalence) and $i$ (rank of the order statistic).

Together with the number of participants $N$, The combination of parameters is constrained by eq. (2.11), which determines the upper bound of $i$ ($i_{max}$) when the other parameters, $\alpha$, and $\gamma_0$ are fixed. Value of $i_{max}$ can find by means of numerical calculation of eq. (2.11) implemented as `check_imax.m`.

Under this constraint, there is no universal optimal set of parameter values because it depends on unknown population characteristics. Here I provide a guideline for selecting the parameters. This is not a solid recommendation but suggested as a first choice. First, $\gamma_0 = 0.5$ may be intuitively acceptable because we can infer that *"a majority (more than half) of the population has label information,"* when we achieve significant result and it is, at present, the standard for existing statistical methods on population prevalence (Friston et al., 1999; see also Discussion in Allefeld et al., 2016). In addition, $\alpha = 0.05$ before the correction for multiple comparisons may be the primary choice because it is the most commonly used in many statistical tests published in neuroscience papers.

As for $i$, if the sufficiently large effect size can be expected and the number of trials is sufficiently large, I recommend to use the largest available $i$ (See Section 3; $i_{max}$). This is because larger $i$ can provide higher statistical power, because *i*-test is tolerant to the existence of $(i-1)$ lower outliers (Section 3). However, if the expected effect size and/or the number of trials is relatively small, the statistical power does not monotonically increase with $i$ (Appendix C). To decide the value in the empirical situation, I provide a MATLAB program `decide_i.m`, which calculates the expected statistical power for each $i$ (Supplementary Figure S4).

<u>Step 2: Identify $\hat{a}_{(i)}$ and estimate $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$</u>

Then, we identify the value of $\hat{a}_{(i)}$ from experimental results. Also, we estimate $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ by empirical test such as permutation test, or assume that it follows a certain distribution, e.g. binomial distribution (Section 2.2.2).

<u>Step 3: Calculate $L$ and compare it with $\alpha$</u>

Finally, by substituting the estimation of $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ to Eq. (2.6) and (2.9), we can calculate the upper bound ($L$) of the probability that the "*i-th order statistic is at*

*or higher than the observed one* ($\hat{a}_{(i)}$)." If $L < \alpha$, we reject the null hypothesis $\mathrm{H}_0: \gamma \leq \gamma_0$ and conclude that $\gamma$ is larger than the predetermined threshold $\gamma_0$ with the significance threshold $\alpha$. The calculation of $L$ is implemented in `itest.m`.

## 3 Numerical calculation from artificial data

We designed *i*-test to improve tolerance of the lower outliers compared to when we fix $i = 1$ (*i*-test-one) as proposed by Allefeld et al., (2016). "Lower outlier" here is defined as a low D-Acc that appears for each participant with small probability. Intuitively, if one lower outlier exists in the experimental results, the order statistic $\hat{a}_{(1)}$ becomes the lower outlier and the statistical test fails to achieve a significant result even when most of the population has brain activity associated with the label ($n \in \Omega_+$). The risk of appearance of one lower outlier becomes larger with the increasing number of participants. Thus, we extended the method so that the second and higher-order statistic can be candidates to make *i*-test tolerant to lower outliers. In this section, I provide an example demonstrating this benefit of *i*-test.

Here, I only provide the numerical calculation results of the probability of the significant result ($P_{sig}$). See Appendix B for detail of the calculation of $P_{sig}$.

Note that although the above discussion implies that the largest $i$ ($i_{max}$) is always best, this is true only when the effect size and the number of trials are sufficiently large (Appendix C).

### 3.1 Definition of artificial experimental results

Here, we consider artificial data, where each participant performed $\mathrm{N}_{trial}$ trials with a binary choice (chance level = 50%). The participants have label information ($n \in \Omega_+$) with a certain probability $\gamma$ and otherwise, do not have label information ($n \in \Omega_-$). We assume that the decoder for participants with label information ($n \in \Omega_+$) predicted the label correctly with probability of $P_{correct+}$(> chance level), while when $n \in \Omega_-$, the decoder for the participant predicted the label in each trial randomly. Also, we assume the independency of each trials. Therefore, the D-Acc for $n \in \Omega_+$ and $n \in \Omega_-$ follows binomial distributions, i.e. $p(a_n|n \in \Omega_+) = BPDF(a_n \times \mathrm{N}_{trial}, \mathrm{N}_{trial}, P_{correct+})$ and $p(a_n|n \in \Omega_-) = BPDF(a_n \times \mathrm{N}_{trial}, \mathrm{N}_{trial}, 0.5)$. Note that strictly speaking, the cross validated decoding accuracies does not perfectly follow binomial distribution in the real fMRI decoding studies, because independency across trials are not fully met (Jamalabadi et al., 2016). In addition, we assume that $P(a_n < \hat{a}_{(i)}|n \in \Omega_-)$ is correctly estimated as $BCDF(\hat{a}_{(i)} \times \mathrm{N}_{trial} - 1, \mathrm{N}_{trial}, 0.5)$.

The threshold for statistical significance is fixed at 0.05 ($\alpha = 0.05$) and the prevalence threshold is fixed at 0.5 ($\gamma_0 = 0.5$). The results are generally replicated with

different values of the prevalence threshold ($\gamma_0 = 0.1,\ 0.3$ and $0.7$; Supplementary Figure S1, S2 and S3).

3.2 Results of numerical calculation

First, I fixed the parameters as follows $\mathrm{N}_{trial} = 1000$, $P_{correct+} = 0.9$ and $\gamma = 0.8$, and I increased the number of participants (N) from 5 to 50 and evaluated the probability of significant results (Figure 2A). Note that when $\mathrm{N} < 5$, the predetermined parameters cannot satisfy the constraint [Eq. (2.11)]; therefore, we cannot apply *i*-test.

The probability of a significant result ($P_{sig}$) of *i*-test-one ($i = 1$) gradually decreases as the number of participants (N) increases from 5 to 50 (thick blue line in Figure 2A). Also, when $i$ is fixed at a certain value (colored thin lines in Figure 2A), the probability of a significant result gradually decreases. However, larger $i$ provided higher $P_{sig}$ and the maximum available value of $i$ ($i_{max}$) also increased with $N$ (Figure 2B). Thus, by using $i_{max}$, *i*-test can improve the statistical power with increase of $N$ (black thick line in Figure 2A). The reduction of D-Acc of *i*-test-one is because the probability of contamination of lower outlier increases when the number of participants increases. In contrast, the improvement of *i*-test with $i_{max}$ is because *i*-test is tolerant to existence of $i - 1$ lower outliers (Appendix B).

Then, we evaluated the effect of $\gamma$ on $P_{sig}$. $\gamma$ was changed from 0 to 1 with 0.01 step (Figure 2C). Note that when $\gamma$ is at or below the prevalence threshold $\gamma_0$ (left side of the Figure 2C), the null hypothesis is true and $P_{sig}$ indicates false positive ratio that should be suppressed below the statistical threshold ($\alpha$). Again, I fixed $\mathrm{N}_{trial} = 1000$, and $P_{correct+} = 0.9$. Figure 2C shows the results from $N = 10, 20, 30, 40, 50$. I found that both *i*-test-one and *i*-test with $i = i_{max}$ could suppress the false alarm ratio below the statistical threshold (Figure 2C left). Also, I found that the statistical power ($P_{sig}$ at $\gamma > \gamma_0$; Figure 2C right) of *i*-test with $i_{max}$ is almost always higher than the *i*-test-one, while the false alarm ratio ($P_{sig}$ at $\gamma \leq \gamma_0$; left side of the figure 2C) of *i*-test with $i_{max}$ is lower for almost all $\gamma \leq \gamma_0$. Exception was observed only when the value of $\gamma$ is close to $\gamma_0$ ($\gamma_0 - 0.04 < \gamma \leq \gamma_0 + 0.01$). This was replicated with the other values of $N$ from 5 to 50 (Results not shown).

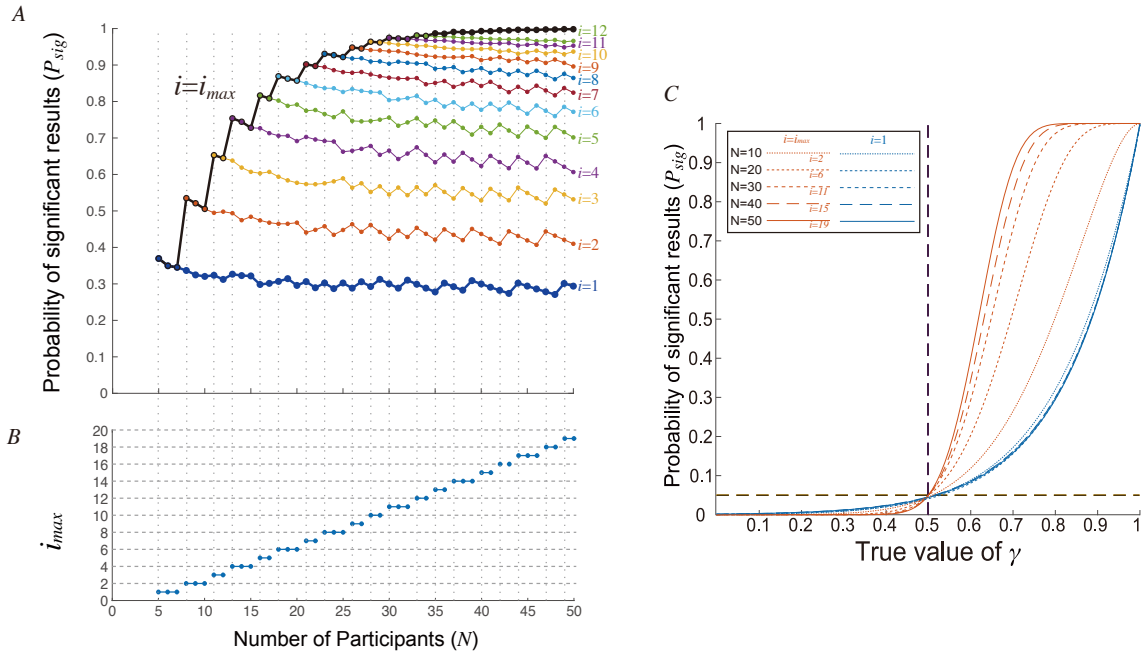



**Figure 2:** Result from the numerical calculation. The fixed parameters for this analysis are $\mathrm{N}_{trial} = 1{,}000$, $P_{correct+} = 0.9$, $\gamma = 0.8$, $\alpha = 0.05$ and $\gamma_0 = 0.5$. **A**: The probability of the significant result ($P_{sig}$) is plotted against the number of participants ($N$). The blue thick line indicates the result of $i$-test-one ($i = 1$). The colored lines indicate the result from fixed $i$ for $i = 1 \ldots 12$. The black thick line indicates the result from $i$-test with a maximum available value of $i$ ($i_{max}$). The vertical black dotted lines indicate the point of $i_{max}$ change. When $N < 5$, the $i$-test is not available because Eq. (2.9) cannot be satisfied with any $i$. **B**: The value of $i_{max}$ against $N$. **C**: The probability of a significant result ($P_{sig}$) are plotted against the value of $\gamma$. Red lines represent results from $i$-test with a maximum available value of $i$ ($i = i_{max}$) and blue lines represent results for $i$-test-one ($i = 1$). The number of participants is 10, 20, 30, 40 or 50 and $i_{max}$ is 2, 6, 11, 15 or 19, respectively. The vertical and horizontal dotted black lines indicate the prevalence threshold $\gamma_0 = 0.5$ and the statistical threshold $\alpha = 0.05$, respectively. When $\gamma > 0.5$ (right side from the vertical dotted line), the alternative hypothesis $\mathrm{H}_1{:}\gamma > \gamma_0$ is true and thus, $P_{sig}$ indicates statistical power and is ideally close to 1, while $\gamma \leq 0.5$, the null hypothesis $\mathrm{H}_0{:}\gamma \leq \gamma_0$ is true and thus, $P_{sig}$ indicates a false alarm ratio and should be smaller than $\alpha$.

## 4 Application to empirical fMRI data

Here, we provide an example of its application to a real fMRI dataset, which is reported in Hirose et al., 2018.

### 4.1 Experiment and analyses procedures

For experimental details, please refer to the original paper (Hirose et al., 2018). Briefly, 12 participants performed a finger-tapping task using either his/her freely-chosen left- or right-hand finger. Each participant underwent 10 sessions, and each session consisted of 15 trials. Each trial started with a 4-sec decision period, where participants chose either their left or right hand to be moved. Then, after a 4-s delay, the participants moved the selected hand finger. Trials were separated by an inter-trial interval (ITI) of 6 s and, in the previous study (Hirose et al., 2018), we carefully evaluated the effect of the previous trial and concluded that there was no carry over of the brain activity for the previous trial. We predicted the chosen hand from brain activity measured just before the movement execution (latter 2 seconds of the delay).

Although each participant underwent 150 trials in total, there were error trials, in which no finger movement occurred, or the participant moved multiple fingers. The number of error trials was 11 for a participant, 5 for another participant, and 1 for other two participants. No error trials were found in the other eight participants. The error trials were excluded from the decoding analysis. Consequently, brain activation for 139 trials for one participant, 145 for another, 149 for other two and 150 for the other eight participants were used for the decoding analysis.

The preprocessed fMRI volume measured before an action initiation is used to predict the chosen hand. The classifier was trained by the iSLR algorithm (Hirose et al., 2015) on voxel activities in the whole brain (31,957 ± 1785). Classification accuracy was evaluated by the leave-one-session-out cross-validation procedure, in which we trained the classifiers using the trials of the nine sessions (training dataset) and used the trials of the remaining session to evaluate the prediction accuracy (test dataset). The balanced accuracy (Brodersen et al., 2010, Brodersen et al., 2012) was calculated from the cross-validation results, because there were bias of the number of trials between labels depending on the bias of the participants' choice.

Also, I performed the permutation test to estimate $p(a_n|n \in \Omega_-)$ with the following procedures. First, the labels (left or right hand) in each session were randomly permuted for each participant. Using this permuted dataset, we then performed the decoding analysis and calculated the decoding accuracy via leave-one-session-out cross validation, using the same procedure as applied to the original (unpermuted) data. To save computational cost, the hyperparameter of the iSLR algorithm (i.e., the number of SLR classifiers) was not re-optimized and the same value as in the analysis without permutation was used in each validation fold. This was repeated 1000 times for each participant, to obtain 1000 decoding accuracies. Then, the decoding accuracies were used to estimate $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ through eq. (2.13). Note that in each permutation, we randomly shuffled the association between brain activities and labels (left or right hand). Therefore, the number of trials assigned with one label is saved in each session.

The preprocessing, first-level decoding analyses, and permutation test for each participant was performed using multi-voxel pattern classification toolbox (https://github.com/satoshi-hirose/MVPC_toolbox).

Then, I applied *i*-test and *i*-test-one to the experimental results and the permutation test results. For all tests, we set the predetermined thresholds as $\alpha = 0.05$ and $\gamma_0 = 0.5$. For the *i*-test, I used $i = 3$, which is the maximum available value that satisfies Eq. (2.11) and was expected to provide highest statistical power when $P_{correct+}$ > 0.65 (Supplementary Figure 4).

## 4.2 Results

*i*-test: The 3rd order statistic was 60.0% ($\hat{a}_{(3)} = 60.0\%$). In the permutation test, the number of permutations with D-Acc lower than $\hat{a}_{(3)}$ was 11,684 (blue bars and green bars in Figure 3) among 12,000 (12 × 1,000) permutations. Therefore, we estimated $P\left(a_n < \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ as 0.97 [11,684/12,000, Eq. (2.13)]. Substituting this value in Eq. (2.6) and Eq. (2.9), we found that $L = 0.024$ and this was smaller than the statistical threshold $\alpha = 0.05$. Thus, the null hypothesis is rejected.

i-test-one: We use the lowest observed D-Acc ($\hat{a}_{(1)} = 47.6\%$) for *i*-test-one. From the permutation results, we found that the D-Acc was lower than 47.6% in 3993 of the 12,000 permutations (green bars in Figure 3). Thus, we estimated $P\left(a < \hat{a}_{(1)} | n \in \Omega_{-}\right)$ as 0.33. Using Eq. (2.6) and (2.16), we found $L = 0.11$. Thus, *i*-test-one could not reject the null hypothesis.

We cannot know the true population for the empirical data. However, when we look at the distribution of the obtained D-Acc (Red bars in Figure 3), the value of $\hat{a}_{(1)}$ can appears to "lower outlier," that unfortunately contaminated the results. Thus, we may assume that *i*-test-one failed to achieve significant results because one participant provides low (outlier) D-Acc. And *i*-test could achieve the significant result, because of its tolerance of lower outliers.

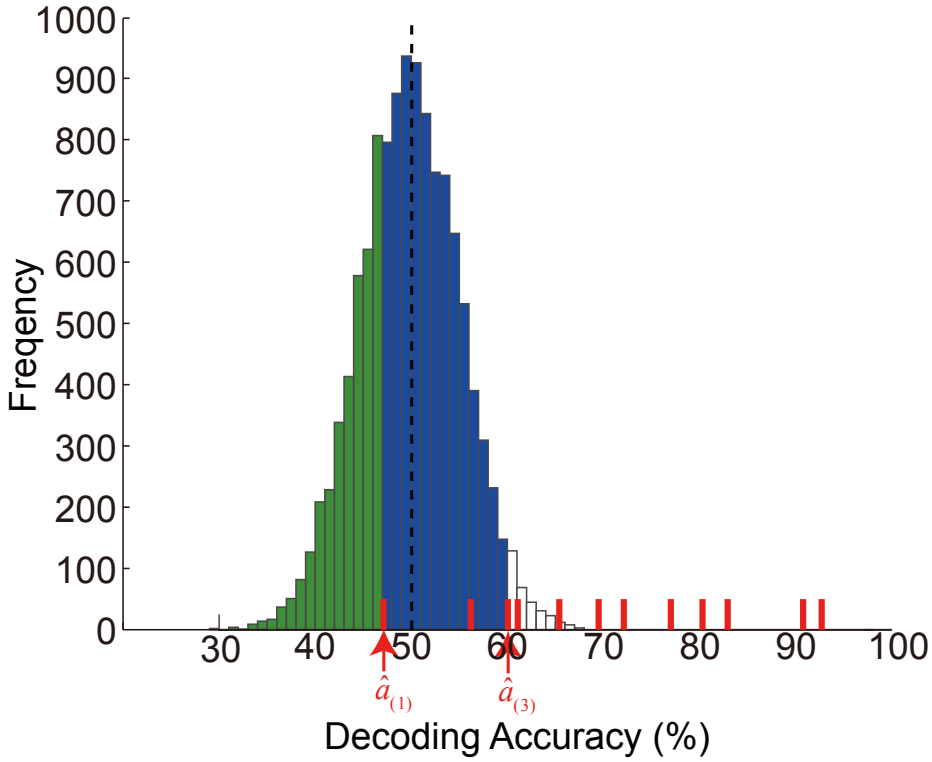



**Figure 3:** Permutation test results in the empirical experiment for the *i*-test with $i = 3$ and i-test-one ($i = 1$). The histogram indicates the frequency of permuted D-Acc $a_{perm(n,m)}$ from 12,000 permutations (12 participants x 1,000 permutations). Green bars indicate the area of $a_{perm(n,m)} < \hat{a}_{(1)}$, whereas $a_{perm(n,m)} < \hat{a}_{(3)}$ is indicated by the blue bars and green bars.

## 5. Discussion

In this study, we proposed a novel statistical test for decoding accuracy and named it *i*-test, which is an extension of the method proposed by Allefeld et al. (2016). In contrast to the commonly used statistical test, which compares the population mean D-Acc with the chance level, our method focuses on population information prevalence, i.e., the proportion of the population that has a higher-than-chance D-Acc. The positive result of the statistical test on population information prevalence can infer major population trends as "*the brain activation has label information in proportion larger than* $\gamma_0$ *of the population (e.g., more than half).*" This is in contrast to the fact that the positive result of the statistical test on the population mean D-Acc can only infer that "*there are some people in the population whose fMRI data carry information about the experimental condition,*" because expectation of each participant's D-Acc should never be lower than the chance level.

A statistical test proposed by Allefeld et al., 2016 (*i*-test-one) also evaluates the population prevalence rather than the population mean. In their method, the minimum statistic is necessarily used for the test. However, as shown in Section 3, using the minimum statistic leads to weakness to lower outliers. This weak point leads to a counter-intuitive drawback: *i*-test-one can reduce its statistical power as the number of participants increases (Figure 2A). This drawback is solved using a higher-order statistic in the *i*-test.

### 5.1 Estimation of $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$

In the analysis of the real experiment (Section 4, Figure 3), I estimated the distribution of D-Acc given that label information is not included in the brain ($p(a_n | n \in \Omega_-)$), by means of the permutation test, in which labels were permuted within each session before the cross validation. However, there are other candidates for the procedures, e.g. permute label among whole experiment and permute only within training/test session for each cross validation. At present, little is known about the advantage and disadvantage of each procedure.

Also, parametric procedure with assumption of binomial distribution (as I have done in the numerical calculation) can be more appropriate in some situation. For

example, when there is limited number of samples, e.g. 4 samples for each label, the maximum number of permutations is 70, which might be insufficient for the empirical estimation of the distribution.

The present study does not cover the validation of these methods. Further study is needed to understand the advantage and disadvantage of the permutation procedures and parametric methods.

## 6. Conclusions

In this study, a novel group-level statistical test for the decoding accuracy named *i*-test is proposed. The advantages of the *i*-test are that the *i*-test 1) can provide a mathematically guaranteed, meaningful population inference; 2) is tolerant to lower outliers and can provide high statistical power.

Although the *i*-test is being introduced as a statistical method for the D-Acc, it is also applicable to other "information-like" measures, such as Mahalanobis distance (Kriegeskorte et al., 2006, Nili et al., 2014), linear discriminant t (Nili et al., 2014) or pattern distinctness D (Allefeld and Haynes, 2014). We expect this study to provide a robust, second-level statistical test for information-based neuroimaging studies.

## 7. Data and code availability

All the analyses in this study can be replicated by MATLAB program codes available at https://github.com/satoshi-hirose/i-test, including implementation of *i*-test and D-Acc obtained from the experiment (Hirose et al., 2018) and from permutation test (1,000 random permutation) for each participant.

**Acknowledgments**

I would like to acknowledge Dr. Atsushi Yokoi, Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, and Dr. Isao Nambu, Nagaoka University of Technology, for helpful suggestions on earlier versions of the manuscript. This work was supported by a Grant-in-Aid for Young Scientists (B) (JSPS KAKENHI 16K16649) to the author.

## Appendix A: Extension for relaxing the assumption of identical distribution among participants

*i*-test can extend to relax the assumption of the identity of the distribution among participants (assumption of i.d.), which is required for most of the existing statistical tests and I used it in *i*-test to derive eq. (2.3). In this extension, we calculate the values of $L$ without this assumption.

Strictly speaking, the assumption of i.d. is often broken in a real fMRI situation. For example, in our experiment (Section 4), the number of trials were different among participants, because of existence of error trials and thus, possible values of decoding accuracy $a_n$ should be different. By considering this issue, at least theoretically the relaxation of the assumption should have a merit to provide a good fit to the true population. However, because we have no clear difference in the empirical and numerical analysis and the relaxation has a demerit on computational cost (Supplementary Figure S5), I only present this as a possible extension of *i*-test.

### A1: Derivation

First, I present the derivation of the extended version. Note that eqs. (2.1), (2.2), (2.4), (2.5), (2.6) and (2.7) is derivation for each participant and therefore, we do not need to replace for the extension.

Here, we do not assume the identical distribution of $p(a_n|n \in \Omega_-)$ among participants, we cannot express "*i-th lowest D-Acc is not less than* $\hat{a}_{(i)}$" with binomial distribution function as eq.(2.3). Instead, $P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$ is expressed with sum of the probability of the all cases;

$$P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right) = \sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} \left(1 - P\left(a_n < \hat{a}_{(i)}|\gamma\right)\right)^{e_n} P\left(a_n < \hat{a}_{(i)}|\gamma\right)^{(1-e_n)} \quad (A.1).$$

Here, $\mathrm{E}_{N,i}$ is the group of all $N$-dimensional vectors, with elements of 1 and 0, and the number of 0s is less than $i$ (the number of 1s is more than $N - i$). When a vector $e$ is an element of $\mathrm{E}_{N,i}$, $e_n$ is the $n$-th element of $e$. The $N$-dimensional vector $e$ represents each combination, and each element of $e$ represents whether the D-Acc of the corresponding participant is higher than $\hat{a}_{(i)}$ (1) or not (0). For example, when $N = 3$ and $i = 2$, $\mathrm{E}_{N,i} = \{[1\ 1\ 1], [0\ 1\ 1], [1\ 0\ 1], [1\ 1\ 0]\}$ and the first element of

$\mathrm{E}_{N,i}$ ([1 1 1]), represents the case that the D-Acc of all three participants are at or higher than $\hat{a}_{(i)}$ and [0 1 1] represents the case that the D-Acc of participant #1 is less than $\hat{a}_{(i)}$ but the remaining two (#2 and #3) are at or higher than $\hat{a}_{(i)}$. The probability of each case is the product of the probability of each participant's result, e.g., when $e =$ [0 1 1], $\prod_{n=1}^{3} P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)^{e_n}\left(1 - P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)\right)^{(1-e_n)} = P\left(a_1 < \hat{a}_{(i)}|\gamma\right)\left(1 - P\left(a_2 < \hat{a}_{(i)}|\gamma\right)\right)\left(1 - P\left(a_3 < \hat{a}_{(i)}|\gamma\right)\right)$.

Then, eqs. (2.8), (2.9) are replaced with

$$\sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} (1 - Q_n)^{e_n} {Q_n}^{(1-e_n)} \geq P\left(a_{(i)} \geq \hat{a}_{(i)}|\mathrm{H}_0 : \gamma \leq \gamma_0\right) \qquad (A.2).$$

$$L = \sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} (1 - Q_n)^{e_n} {Q_n}^{(1-e_n)} \qquad (A.3).$$

A notice for the derivation of A.2 is that $P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$ in Eq. (A.1) monotonically decreases with each $P\left(a_n < \hat{a}_{(i)}|\gamma\right)$. The monotonicity is obvious because when the probability that the "*D-Acc of one of the participants is less than* $\hat{a}_{(i)}$" ($P\left(a_n < \hat{a}_{(i)}|\gamma\right)$) increases, the probability that "*more than* $N - i$ *participants have a D-Acc equal to or higher than* $\hat{a}_{(i)}$" ($P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$) should decrease. Then, by comparing $L$ with $\alpha$, we can perform *i*-test without the assumption of identical distribution among participants.

*A.2 Empirical example*

Here, I applied the extended *i*-test to the experimental results used in Section 4 (Hirose et al., 2018). As in Section 4, we set the predetermined parameters as $\alpha = 0.05$, $\gamma_0 = 0.5$. and $i = 3$. Unlike the original *i*-test, the estimation of the D-Acc distribution given that a participant does not have label information in the brain ($p(a_n|n \in \Omega_-)$) should be done for each participant. Thus, eq. (2.12) is replaced with the following equation for each participant;

$$P\left(a_n < \hat{a}_{(i)}|n \in \Omega_-\right) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{a}_{perm(n,m)} < \hat{a}_{(i)}\right] \qquad (A.4).$$

Here, we found that the estimated values ranged from 0.963 to 0.983. By substituting these values into Eqs. (2.6) and (A.3), we found $L = 0.024$, which is similar to the original *i*-test (difference < 0.000002). Thus, this extended version also rejected the null hypothesis.

*2.2 Computational cost*

A demerit of this extension is its large computational cost. Because elements of $\mathrm{E}_{N,i}$ rapidly increase with $N$ and $i$, the calculation of A.3 easily suffers a combination explosion of computational cost. Empirical benchmark test identified that the computation of the extended *i*-test with $i_{max}$ was completed within 1 s when $N < 23$ and was completed within 1 min when $N < 28$, but it took about 5.5 min for $N = 30$ (Supplementary Figure 6A). Thus, the empirical upper bound of $N$ may be about 30. This is contrasted with the original *i*-test that requires computational time that linearly increases with $N$ and requires less than 0.03 s for $N = 1{,}000$ (Supplementary Figure 6B).

## Appendix B: Numerical calculation of $P_{sig}$

Here I describe the calculation of the probability of the significant result ($P_{sig}$) reported in Section 3. The outline of the calculation is as follows. First, I introduce the threshold of the $i$-th order statistic; $T$, such that $i$-test provides positive results when and only when $a_{(i)} > T$ (Section B1). The value of T can be derived from the true DA distribution (γ, $p(a_n|n \in \Omega_-)$ and $p(a_n|n \in \Omega_+)$) and the predetermined parameters ($\alpha$, $\gamma_0$ and $i$). Note that here, I assume the binomial distribution for $p(a_n|n \in \Omega_-)$ and $p(a_n|n \in \Omega_+)$. Then, I calculate the probability that $i$-th order statistic is larger than $T$ ($P(a_{(i)} > T)$; Section B2).

To improve understanding, numerical example with the following parameters are provided in each step; $\alpha = 0.05, \gamma_0 = 0.5$, $N = 50$, $i$=1, $N_{trial} = 1{,}000$, $P_{correct+} = 0.9$ and $\gamma = 0.8$.

### B1 Introduction of the threshold of $i$-th order statistic; $T$

From the assumptions, $P(a_n < a_{(i)}|n \in \Omega_-)$ is expressed with binomial cumulative distribution function as follows;

$$P(a_n < a_{(i)}|n \in \Omega_-) = BCDF(a_{(i)} \times \mathrm{N}_{trial} - 1, \mathrm{N}_{trial}, 0.5) \qquad (B.1).$$

Here, I repeat the equation for $i$-test [i.e., Eq. (2.6) and Eq. (2.9)].

$$Q_n = (1 - \gamma_0)P(a_n < a_{(i)}|n \in \Omega_-) \qquad (2.6),$$

$$L = BCDF(i - 1, N, Q_n) \qquad (2.9).$$

Note that here, the $i$-th order statistic is not a constant but probabilistic variable and therefore, it is expressed without hat. Form eqs. (B.1), (2.6) and (2.9), $L$ can be regarded as a (discrete) function of $a_{(i)}$. $L$ decreases with $\hat{a}_{(i)}$, because $P(a_n < \hat{a}_{(i)}|n \in \Omega_-)$ increases with $a_{(i)}$, $Q_n$ increases with $P(a_n < \hat{a}_{(i)}|n \in \Omega_-)$ and $L$ decreases with $Q_n$. Therefore, we can define the threshold $T$, such that $\alpha > L$

when and only when $\hat{a}_{(i)} > T$. Because the D-Acc is discrete, we can easily perform numeric calculation of $T$ from $p(a_n|n \in \Omega_-)$, $\gamma_0$, $i$, $N$ and $N_{trial}$, without analytical solution (Supplementary Figure S6).

In the example case with parameters listed above we find $T = 0.481$.

<u>B2 Calculate the probability that $i$-th order statistic is higher than $T$</u>

By definition, the probability of the significant result is determined as follows:

$$P_{sig} = P(\alpha > L) \qquad (B.2).$$

From equivalence of $\alpha > L$ and $a_{(i)} > T$, we find

$$P_{sig} = P(a_{(i)} > T) \qquad (B.3).$$

Below, the goal is to calculate $P(a_{(i)} > T)$. First, I formulate the probability that "*the D-Acc of a participant is at or less than $T$,*" as follows;

$$P(a_n \leq T) = \gamma P(a_n \leq T|n \in \Omega_+) + (1-\gamma)P(a_n \leq T|n \in \Omega_-) \qquad (B.4).$$

Then, by noting that "*$i$-th order statistic is higher than the threshold* ($T$)" is identical to "*fewer than $i$ among $N$ participants have a D-Acc at or lower than $T$,*" $P(\hat{a}_{(i)} > T)$ can be expressed as follows;

$$P(a_{(i)} > T) = BCDF(i-1, N, P(a_n \leq T)) \qquad (B.5).$$

From eq.(B.4) together with true distribution ($p(a_n|n \in \Omega_+)$, $p(a_n|n \in \Omega_-)$ and $\gamma$), we can calculate the value of $P(a_n \leq T)$. Then, from eq. (B.5), we can calculate $P(a_{(i)} > T)$, which is equal to $P_{sig}$.

In the example, $P(a_n \leq T|n \in \Omega_+)$ and $P(a_n \leq T|n \in \Omega_-)$ was, as follows;

$$P(a_n \leq T|\Omega_+) = BCDF(T \times \mathrm{N}_{trial}, \mathrm{N}_{trial}, P_{correct+}) = 1.43 \times 10^{-242}$$

$$P(a_n \leq T|\Omega_-) = BCDF(T \times \mathrm{N}_{trial}, \mathrm{N}_{trial}, 0.5) = 0.121$$

Then, substituting them together with $\gamma = 0.8$ into eq.(B.4), we find $P(a_n \leq T) = 0.0242$. By substituting it to eq.(B.5) together with $i = 1$ and $N = 50$, we find $P(\hat{a}_{(i)} > T) = 0.294$.

This means that, although the probability that "*D-Acc of a participant is at or less than* $T$" $P(a_n \leq T)$ is small (3%), the probability that *"D-Acc of <u>no</u> participant among* $N$ *is less than* $T$" $P(\hat{a}_{(1)} > T)$ is only 29%, because there are many (50) participants. As a result, *i*-test-one failed to achieve significant results with probability of more than 70%, although the true $\gamma$ (0.8) is much larger than the prevalence threshold $\gamma_0 = 0.5$.

This is solved by using larger $i$. When we use the parameter as $i = 19$ ($i_{max}$), the threshold T becomes slightly higher ($T = 0.530$) and as a results, the probability that "*the D-Acc of a participant is less than* $T$*,"* becomes larger $P(a_n \leq T) = 0.195$. However, with $i = 19$, *i*-test can provide significant result when *"D-Acc of fewer than 19 participants among* $N$ *is less than* $T$*."* This indicates that *i*-test can be tolerant to the 18 "lower outliers" (i.e. participants with D-Acc lower than $T$) and as a result, *i*-test can provide significant result almost 100% ($P(\hat{a}_{(19)} > T) = 0.998$). These results can be linked with my design concept of *i*-test for the tolerance of lower outlier. Namely, when we define D-Acc below the threshold $T$ as "lower outlier," the tolerance of lower outlier can be achieved by generalize the *i*-test-one (Allefeld et al., 2016) to *i*-test and as a result, *i*-test can achieve higher statistical power.

## Appendix C: $i_{max}$ is not always optimal

As demonstrated in the Section 3, larger *i* leads tolerance of lower outlier and consequently can provide higher statistical power. This implies that the largest $i$ ($i_{max}$) is the optimal choice. However, this is not a general fact.

For example, when the probability of the correct decoding for $\Omega_+$ ($P_{correct+}$) is low ($P_{correct+} = 0.53$), the relation between $i$ and $P_{sig}$ is not monotonical but inverse-U shape (Appendix Figure C.1A). Another example is shown in Figure 3B. When the number of trials is small ($\mathrm{N}_{trial} = 10$), the relation seems to be more complex; $P_{sig}$ generally increased with $i$ but, decreases when $i$ increases from 1 to 2, from 5 to 6, from 10 to 11, from 15 to 16 and from 18 to 19 (Appendix Figure C.1B).

These phenomena can be explained as follows. From the derivation of $P_{sig}$ (Appendix B), we can find that increase of $i$ affect $P_{sig}$ in two pathways. First (increase effect), when $i$ increases, the number of ignorable lower outliers ($a_n \leq T$) increases (eq.(B.5)). Therefore, it increases the statistical power as demonstrated in Section 3. Second (decrease effect), the threshold $T$ can increase as $i$ increases and consequently, probability of "lower outlier" $P(a_n \leq T)$ can increase. Then, increase of $P(a_n \leq T)$ leads decrease of $P_{sig}$. The decrease effect becomes dominant when $i$ is large in the first example with low $P_{correct+}$. And in the second example, decrease effect appears only at certain points, because of the discreteness of possible values of $T$.

Anyhow, the optimal $i$ is decided with the balance between these two effects and the effects depends on the known parameters, such as number of trials, as well as unknown true parameters, such as the probability of the correct decoding for participants belonging to $\Omega_+$ ($P_{correct+}$).

However, the parameter $i$ should be decided without observation of data to avoid inflation of false positive rate by "cheating effect." For the purpose, I recommend selecting $i$ that maximizes expected statistical power as usually done in "a priori power analysis" (Faul et al., 2009, G*power; http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html). For this parameter determination, I provide a program code (decide_i.m) that enable us to calculate the statistical power under the assumption

of binomial distribution of the decoding accuracy. An example is shown in Supplementary Figure 4.

Reference

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior research methods, 41(4), 1149-1160.

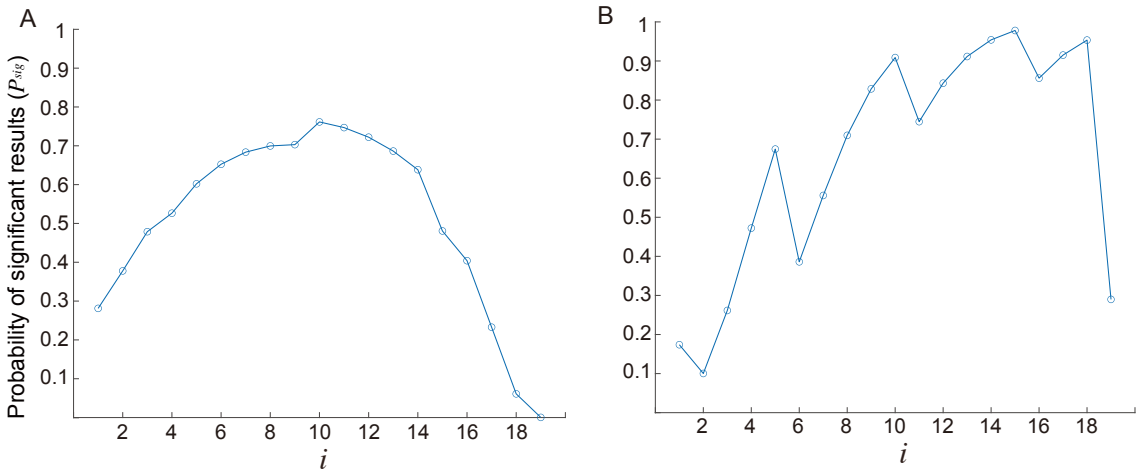



Figure A1: Result from the numerical calculation. **A**: Effect of low effect size ($P_{correct+} = 0.53$). The probability of the significant result ($P_{sig}$) is plotted against the parameter $i$. The number of participants was fixed as $\mathrm{N} = 50$. The other fixed parameters were the same as those of Section 3.2 ($\mathrm{N}_{trial} = 1{,}000$, $\gamma = 0.8$, $\alpha = 0.05$ and $\gamma_0 = 0.5$). **B**: Effect of small number of trials ($\mathrm{N}_{trial} = 10$). The probability of the significant result ($P_{sig}$) is plotted against the parameter $i$. The number of participants was fixed as $\mathrm{N} = 50$. The other fixed parameters were same as those of Section 3.2 ($\gamma = 0.8$, $P_{correct+} = 0.9$, $\alpha = 0.05$ and $\gamma_0 = 0.5$).

## References


Allefeld, C., Görgen, K., Haynes, J.D., 2016. Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. Neuroimage. 141, 378–392. DOI: 10.1016/j.neuroimage.2016.07.040

Allefeld, C., Haynes, J.D. 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage. 89, 345-357. DOI : 10.1016/j.neuroimage.2013.11.043

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M. The balanced accuracy and its posterior distribution. In Pattern Recognition (ICPR), 2010 20th international conference on (pp. 3121–3124). IEEE. DOI: 10.1109/ICPR.2010.764

Brodersen, K.H., Mathys, C., Chumbley, J.R., Daunizeau, J., Ong, C.S., Buhmann, J.M., Stephan, K.E., 2012. Bayesian mixed-effects inference on classification performance in hierarchical data sets. J. Mach. Learn. Res. 13, 3133–3176.

Deuker, L., Olligs, J., Fell, J., Kranz, T.A., Mormann, F., Montag, C., Reuter, M., Elger, C.E., Axmacher, N., 2013. Memory consolidation by replay of stimulus-specific neural activity. J. Neurosci. 33(49), 19373–19383. DOI: 10.1523/JNEUROSCI.0414-13.2013

Etzel, J.A., Zacks, J.M., Braver, T.S. 2013. Searchlight analysis: promise, pitfalls, and potential. Neuroimage. 78, 261–269. DOI: 10.1016/j.neuroimage.2016.07.040

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S., Turner, R., 1995. Analysis of fMRI time-series revisited. Neuroimage. 2(1), 45–53. DOI: 10.1006/nimg.1995.1023

Friston, K.J., Holmes, A.P., Price, C.J., Büchel, C., Worsley, K.J., 1999. Multisubject fMRI studies and conjunction analyses. Neuroimage. 10(4), 385–396. DOI: 10.1006/nimg.1999.0484

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., Düzel, E., 2010. Theta-coupled periodic replay in working memory. Curr. Biol. 20(7), 606–612. DOI: 10.1016/j.cub.2010.01.057

Gallivan, J.P., McLean, D.A., Valyear, K.F., Pettypiece, C.E., Culham, J.C., 2011. Decoding action intentions from preparatory brain activity in human parieto-frontal networks. J. Neurosci. 31, 9599–9610. DOI: 10.1523/JNEUROSCI.0080-11.2011

Gallivan, J.P., McLean, D.A., Flanagan, J.R., Culham, J.C., 2013. Where one hand meets the other: limb-specific and action-dependent movement plans decoded from preparatory signals in single human frontoparietal brain areas. J. Neurosci. 33, 1991–2008. DOI: /10.1523/JNEUROSCI.0541-12.2013

Gilbert, S.J., Fung, H., 2018. Decoding intentions of self and others from fMRI activity patterns. Neuroimage. 172, 278-290. DOI: 10.1016/j.neuroimage.2017.12.090

Good, P., 2000. Permutation tests: A practical guide to resampling methods for testing hypotheses. Springer

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456. DOI: 10.1146/annurev-neuro-062012-170325

Hirose, S., Nambu, I., Naito, E., 2015. An empirical solution for over-pruning with a novel ensemble-learning method for fMRI decoding. J. Neurosci. Methods. 239, 238–245. DOI: 0.1016/j.jneumeth.2014.10.023

Hirose, S., Nambu, I., Naito, E., 2018. Cortical activation associated with motor preparation can be used to predict the freely chosen effector of an upcoming movement and reflects response time: An fMRI decoding study. Neuroimage. 183, 584–596. DOI: 10.1016/j.neuroimage.2018.08.060

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., Gais, S., 2016. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum. Brain. Mapp. 37(5), 1842–1855. DOI: 10.1002/hbm.23140

Kriegeskorte, N., Goebel, R., Bandettini, P. 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103(10), 3863-3868. DOI: 10.1073/pnas.0600244103

Kowalczyk, A. Classification of anti-learnable biological and synthetic data. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 176–187). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-74976-9_19

Kowalczyk, A., Greenawalt, D.M., Bedo, J., Duong, C., Raskutti, G., Thomas, R.J., Phillips, W.A., 2007. Large Validation of Anti-learnable Signature in Classification of Response to Chemoradiotherapy in Esophageal Adenocarcinoma Patients. The First International Symposium on Optimization and Systems Biology (OSB'07).

Nambu, I., Hagura, N., Hirose, S., Wada, Y., Kawato, M., Naito, E., 2015. Decoding sequential finger movements from preparatory activity in higher-order motor regions: a functional magnetic resonance imaging multi-voxel pattern analysis. Eur. J. Neurosci. 42, 2851–2859. DOI: 10.1111/ejn.13063

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N. 2014. A toolbox for representational similarity analysis. PLoS Comput Biol. 10(4), e1003553. DOI: 10.1371/journal.pcbi.1003553

Nishida, S., Nishimoto, S., 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage. 180, 232–242. DOI: 10.1016/j.neuroimage.2017.08.017

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. Neuroimage Clin. 4, 687–694. DOI: 10.1016/j.nicl.2014.04.004

Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res. 11(Jun), 1833–1863.

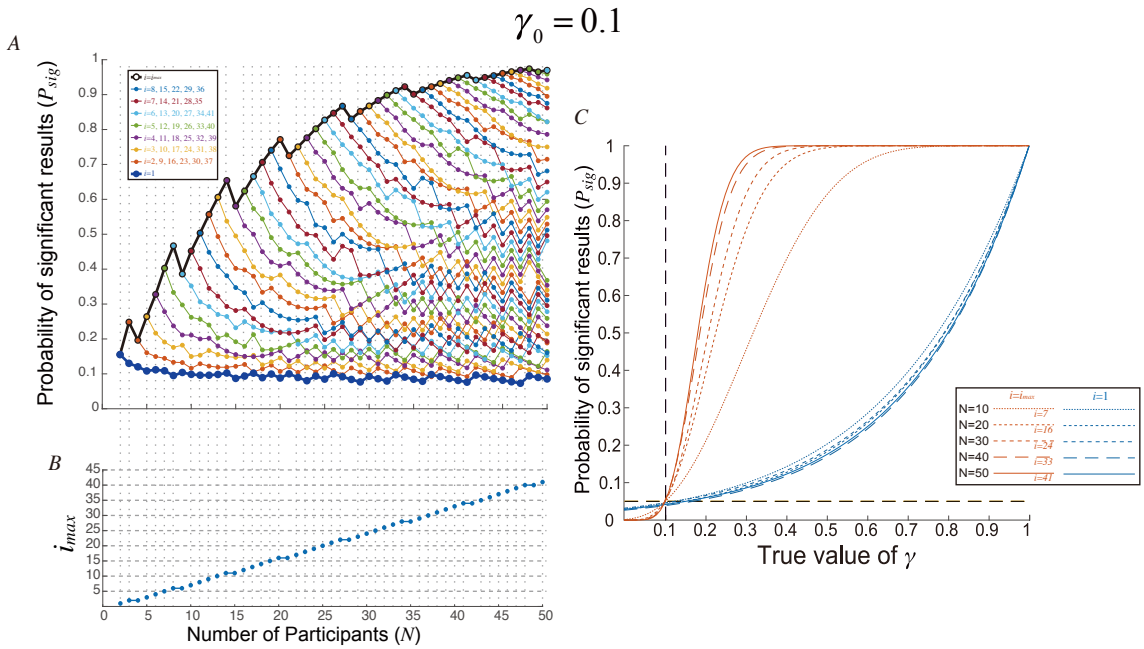
Roadknight, C., Aickelin, U., Qiu, G., Scholefield, J., Durrant, L. Supervised learning and anti-learning of colorectal cancer classes and survival rates from cellular biology parameters. In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference. DOI: 10.1109/ICSMC.2012.6377825

Rosenblatt, J.D., Vink, M., Benjamini, Y., 2014. Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. Neuroimage. 84, 113–121. DOI: 10.1016/j.neuroimage.2013.08.025

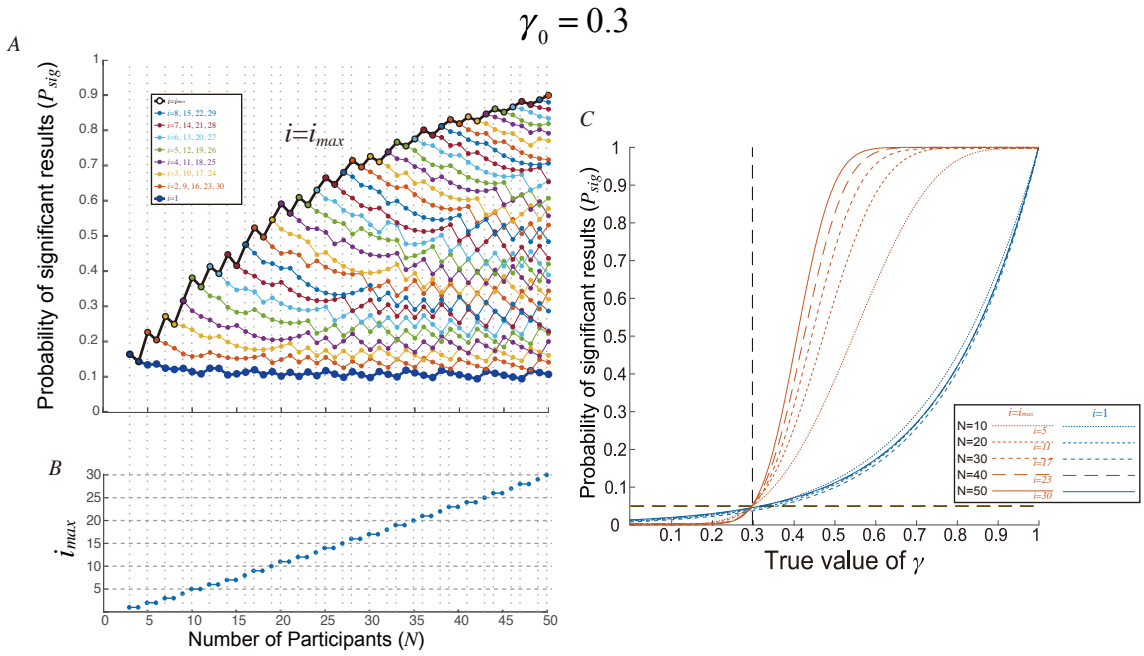
Spiridon, M., Kanwisher, N., 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron. 35(6), 1157–1165. DOI: 10.1016/S0896-6273(02)00877-2

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage. 65, 69–82. DOI: 10.1016/j.neuroimage.2012.09.063

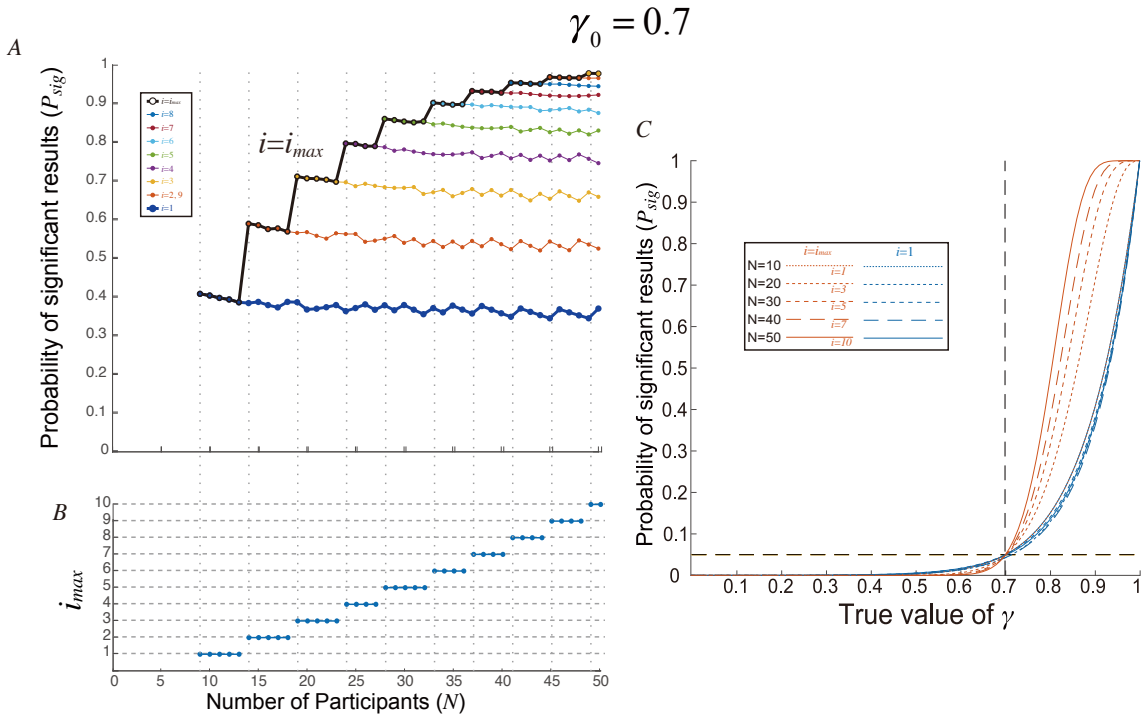
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage. 46 (4), 1004-1017. DOI: 10.1016/j.neuroimage.2009.03.025



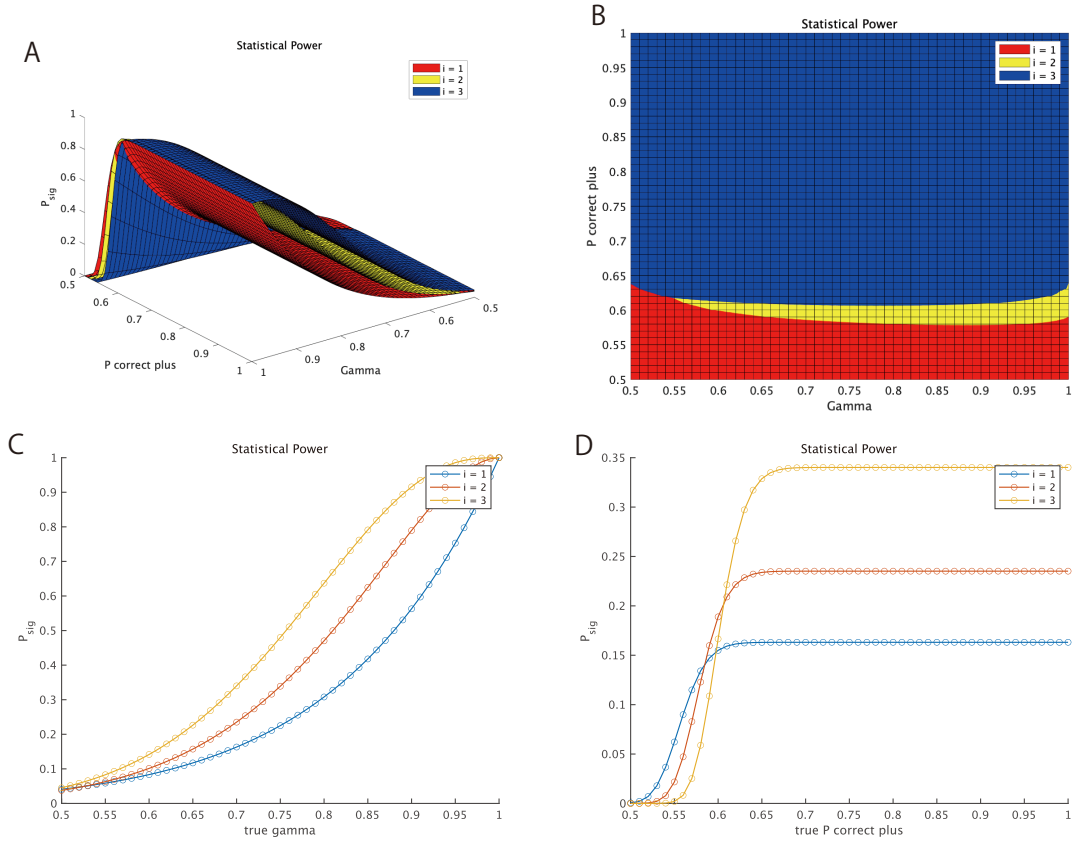

Supplementary Figure S1: Results from the numerical calculation with $\gamma_0 = 0.1$ is shown following the identical format to Figure 2. True gamma is set to 0.3. The other parameters are identical to those reported in the main text, i.e. $N_{trial} = 1{,}000$, $P_{correct+} = 0.9$, $\alpha = 0.05$.



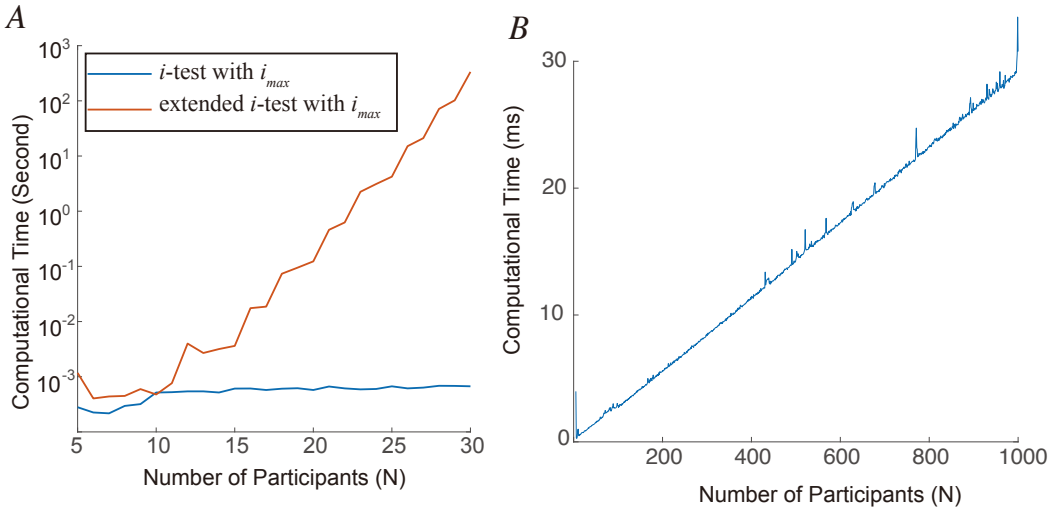

Supplementary Figure S2: Results from the numerical calculation with $\gamma_0 = 0.3$ is shown following the identical format to Figure 2. True gamma is set to 0.5. The other parameters are identical to those reported in the main text, i.e. $N_{trial} = 1{,}000$, $P_{correct+} = 0.9$, $\alpha = 0.05$.



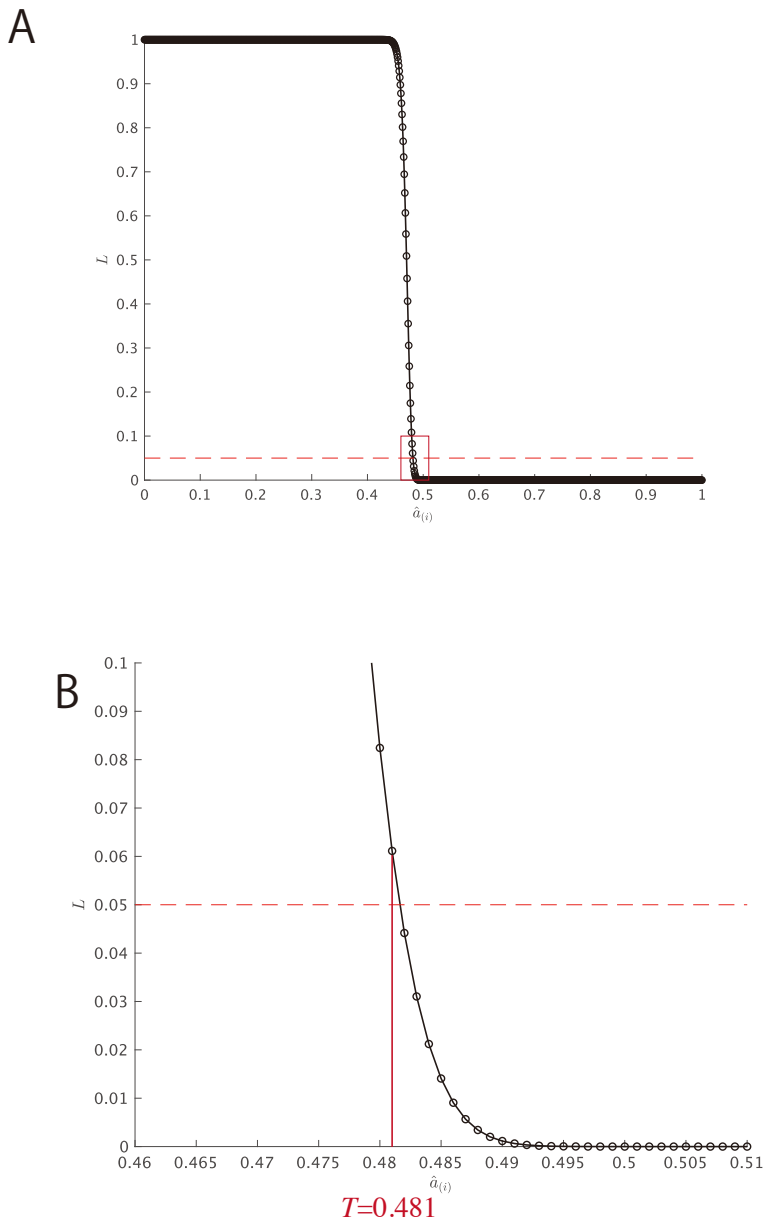

Supplementary Figure S3: Results from the numerical calculation with $\gamma_0 = 0.7$ is shown following the identical format to Figure 2. True gamma is set to 0.9. The other parameters are identical to those reported in the main text, i.e. $N_{trial} = 1{,}000$, $P_{correct+} = 0.9$, $\alpha = 0.05$.




Supplementary Figure S4: Statistical power expectation for the decision of $i$ in the empirical experiment ($\mathrm{N} = 12$, $\mathrm{N}_{trial} = 150$). I set the predetermined threshold parameters as $\gamma_0 = 0.5$ and $\alpha = 0.05$. A: Expected statistical power ($P_{sig}$; z-axis) is plotted against $P_{correct+}$ (x-axis, from 0.5 to 1 with 0.01 steps) and $\gamma$ (y-axis, from 0.5 to 1 with 0.01 steps) for $i = 1, i = 2$ and $i = 3$ ($i_{max}$). B: Right from above view. Red, yellow and blue regions indicate that $i = 1$, $2$ and $3$ provides maximum statistical power. C: $P_{correct+}$ is fixed at 0.7 and expected statistical power ($P_{sig}$; y-axis) is plotted against $\gamma$ (x-axis). D: $\gamma$ is fixed at 0.7 and expected statistical power ($P_{sig}$; y-axis) is plotted against $P_{correct+}$ (x-axis). All panels are illustrated by the MATLAB program "`decide_i.m`."




Supplementary Figure S5: Benchmark test results on MATLAB 2018a, 64 bit, Linux, with 4 core of 3.5GHz CPU (Intel Xeon). **A**: Computational time for original i-test with maximum $i$ (blue) and the extended i-test without identical distribution assumption (Appendix A; red) is plotted against the number of participants $N$. The vertical axis is log-scaled. **B**: Computational time for i-test with maximum $i$. The vertical axis is linear scaled.




Supplementary Figure S6: Calculation of the threshold $T$. The parameters are as follows; $\alpha = 0.05, \gamma_0 = 0.5$, $N = 50$, $i$=1, $N_{trial} = 1{,}000$, $P_{correct+} = 0.9$ and $\gamma = 0.8$. A: $L$ is plotted against each value of $a_{(i)}$. Note that $a_{(i)}$ can take discrete values; from 0 to 1 with 0.001 steps. Each black circle corresponds to each value of $a_{(i)}$. Horizontal dotted line indicates the statistical threshold $\alpha = 0.05$. B: Zoom-up view of the region indicated by the red square in Panel A. Horizontal dotted line indicates the statistical threshold $\alpha = 0.05$. Results indicate that $L$ is smaller than $\alpha$ when and only when $a_{(i)}$ is larger than 0.481 ($T = 0.481$).