False Discovery Rate Control for Lesion Symptom Mapping with Heterogeneous data via Weighted P-values

Siyu Zheng[a], Alexander C. McLain[a*], Joshua Habiger[b],
Christopher Rorden[c] and Julius Fridriksson[d]

[a]Department of Epidemiology and Biostatistics,
University of South Carolina.
[b]Department of Statistics, Oklahoma State University
[c]Department of Psychology, University of South Carolina
[d]Department of Communication Sciences and Disorders,
University of South Carolina

August 17, 2023

Abstract

Lesion-symptom mapping studies provide insight into what areas of the brain are involved in different aspects of cognition. This is commonly done via behavioral testing in patients with a naturally occurring brain injury or lesions (e.g., strokes or brain tumors). This results in high-dimensional observational data where lesion status (present/absent) is non-uniformly distributed with some voxels having lesions in very few (or no) subjects. In this situation, mass univariate hypothesis tests have severe power heterogeneity where many tests are known *a priori* to have little to no power. Recent advancements in multiple testing methodologies allow researchers to weigh hypotheses according to side-information (e.g., information on power heterogeneity). In this paper, we propose the use of p-value weighting for voxel-based lesion-symptom mapping (VLSM) studies. The weights are created using the distribution of lesion status and spatial information to estimate different non-null prior probabilities for each hypothesis test through some common approaches. We provide a *monotone minimum weight* criterion which requires minimum *a priori* power information. Our methods are demonstrated on dependent simulated data and an aphasia study investigating which regions of the brain are associated with the severity

*email: mclaina@mailbox.sc.edu

of language impairment among stroke survivors. The results demonstrate that the proposed methods have robust error control and can increase power. Further, we showcase how weights can be used to identify regions that are inconclusive due to lack of power.

Keywords: Heterogeneous data; False discovery rate; Neuroimaging data; Voxel-based lesion symptom mapping; Weighted p-values.

1   Introduction

Data arising from neuroscience studies have considerable statistical issues including a large number of parameters, an unknown spatial dependence structure, and (commonly) low statistical power. Neuroimaging consists of using magnetic resonance imaging (MRI), positron emission tomography (PET), electroencephalography (EEG), or other imaging modalities, to measure various aspects of brain structure and activity. Data modalities from MRI include functional MRI (fMRI), structural T1 weighted images (T1), and diffusion-weighted imaging (DWI) among others. These data are typically measured on a voxel level in three-dimensional space. As imaging technologies improve the number of data voxels per scan has increased, possibly reaching into the millions depending on the spatial resolution of the image. Independent statistical tests are often computed for each location. Therefore, as spatial resolution increases, the opportunity for making erroneous discoveries increases. This requires some principled thresholding to control for global type I error rate at a level α. Common criteria on the global type I error rate include the familywise error rate (FWER) (Tukey, 1994; Nichols and Holmes, 2002) and the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

Recent statistical methodology has considered using prior information about the hypotheses to improve results through p-value weighting (Genovese et al., 2006; Roeder and Wasserman, 2009; Peña et al., 2011; Habiger, 2017; Ignatiadis and Huber, 2021; Lei and Fithian, 2018; Zhang and Chen, 2020), grouping similar hypotheses (Cai and Sun, 2009; Hu et al., 2010; Ignatiadis et al., 2016), or weighting global type I error rate criteria (Benjamini and Hochberg, 1997; Benjamini and Cohen, 2017; Basu et al., 2018). P-value weighting is

a procedure that uses prior information on hypotheses heterogeneity to improve the overall power. This prior information – commonly referred to as *side-information* – can consist of results from previous studies on the most 'promising' hypotheses (Li and Barber, 2017, 2019; Lei and Fithian, 2018), covariate data indirectly related to the hypotheses (Ignatiadis and Huber, 2021), or information related to the heterogeneity in the power functions of the hypotheses (Peña et al., 2011; Habiger, 2017). The goal of a p-value weighting procedure is to design the weights to maximize the expected number of discoveries while controlling the FWER or FDR.

Modern weighting methods commonly use regression-type approaches to incorporate the side-information into the multiple testing procedure. Commonly these methods use the conditional two-group model where the side-information impacts the probability a test is null and the non-null p-value distribution. For example, Lei and Fithian (2018) proposed adaptive p-value thresholding (AdaPT) which adaptively estimates a Bayes optimal p-value rejection threshold. This is done through the use of the Expectation Maximization (EM) algorithm using a set of partially masked p-values. A similar approach referred to as covariate adaptive multiple testing (CAMT) by Zhang and Chen (2020), also uses the EM algorithm with their M step being expressed in terms of the ratio of alternative and null distributions which is modeled using the beta density. Ignatiadis and Huber (2021) proposed Independent Hypothesis Weighting (IHW) which divides all tests into several independent folds. For each fold, the estimated weight function can be learned from the p-values and covariates in the remaining folds. Similar to the AdaPT, IHW estimates the null probability and non-null distribution based on a conditional two-group model via an EM algorithm. AdaPT and IHW have been shown to provide finite sample FDR control, while CAMT can provide asymptotic FDR control. Cai et al. (2021) proposed a locally adaptive weighting and screening (LAWS) method to deal with spatial multiple testing problems. The LAWS procedure estimates the weights by using the spatial structure through a kernel screening method and can control the FDR asymptotically. Boca and Leek (2018) proposed an FDR control multiple testing method (R package `swfdr` Leek et al., 2021) where – in

the spirit of the Storey (2002) procedure – the unknown null indicator is replaced with an indicator the p-value is lower than some threshold. The indicators are used to estimate their associations with the side-information.

In this paper, we expand the Weighted Adaptive Benjamini Hochberg (WABH) procedure proposed by Habiger (2017) to incorporate heterogeneous non-null probabilities and effect sizes. Further, we demonstrate how our methods are flexible to specific statistical models and are tailored to perform well in low-power settings, which are common in our application to voxel-based lesion-symptom mapping (VLSM) analyses (Bates et al., 2003; Rorden et al., 2007). Further, WABH is known to be robust to poor estimation of the parameters governing the impact of the side-information on the weights. No previous methods are available that are designed for low-power settings and are robust to misspecification of the weights. In general, this procedure can be applied to any situation where the p-values arise from mass univariate logistic regression. Below, we detail the statistical issues that arise within VLSM and the solutions provided by our testing procedure.

1.1   Voxel-based lesion symptom mapping

A main goal of research in neuroscience is to identify and examine areas of the brain related to behavioral or cognitive functions. A common method is to use subjects with a recent brain injury (e.g., from a traumatic brain injury, epilepsy, or stroke) to map some domain of cognition to specific regions of the brain. This can provide theoretical insights regarding brain function and can also inform clinical treatment. The most popular lesion-symptom mapping approach – voxel-based lesion-symptom mapping (VLSM) (Bates et al., 2003; Rorden et al., 2007)– typically relies on structural MRI images (e.g., fMRI, T1, or DWI) where lesion status is measured on parcellated three-dimensional voxels (e.g. 1 mm^3) and relates lesion status to an outcome of interest in each voxel (see Karnath et al., 2018, for a recent review of the field). The number of tests in a VLSM can reach millions depending on the resolution of the brain scan.

VLSM analyses are typically mass-univariate tests consisting of computing a simple test

statistic (e.g., t-tests, General Linear Models, etc.) independently for each voxel and then using some multiple testing correction to identify significant associations with regions of the brain. There are a number of statistical issues that complicate such analyses. First, since studies in humans cannot be designed to injure certain areas of the brain we must rely on naturally occurring injuries (Rorden and Karnath, 2004). This commonly results in lesions being unequally distributed, with some areas/voxels having lesions in a few subjects. For example, stroke-related brain injury is determined by vasculature leading to some regions being far more vulnerable than others. Therefore, the spatial sampling of lesions is not random, and statistical power varies across space. Consider a study of language impairment following left hemisphere injury: we will have low power in regions typically spared in stroke and no power in the right hemisphere (as we have no variability). In response, some have advocated only using voxels that are impacted in, for example, 10% of subjects to account for this issue (Holmes et al., 1996). Second, it is likely that the areas of the brain that will impact the cognitive outcome will have some spatial clustering. That is, the 3-dimensional coordinates of the voxels will be related to the non-null probability. Ignoring this relationship misses out on an important source of variation in the signal, thereby decreasing the overall power of the procedure.

The above issues naturally fit into the purview of weighted multiple testing. The naturally occurring injuries in VLSM create heterogeneous power among the voxels. Voxel power is a function of effect size, which is commonly unknown in practice. In this paper, we propose and provide a straightforward solution consisting of a *monotone minimum weight* criterion, which automatically estimates voxel power and has desirable properties for studies with low to moderate power. Further, we test using plug-in estimates of the non-null probability using state-of-the-art methods (AdaPT and CAMT) which utilize any spatial clustering of signals to gain power. In our data analysis, we demonstrate how the presentation of the impact of weighting is key for transparent reporting of weighted analyses.

The outline of the paper is as follows. In Section 2, we review multiple testing pro-

cedures for data with heterogeneity among the hypotheses. In Section 3, we discuss how weighted multiple testing procedures can be applied to VLSM in a number of common scenarios. In Section 4, we present results from numerous simulation studies that compare the performance of the proposed method to some common approaches. In Section 5, we present an analysis of 220 individuals with chronic left hemisphere stroke and identify areas of the brain associated with the severity of aphasia, a language disorder that impacts the expression and comprehension of speech.

2 Multiple Testing with Heterogeneous Data

2.1 Setup and Notation

Consider testing null hypothesis H_m based on the random vector $\mathcal{D}_m$ for $m = 1, 2, \ldots, M$. The decision to reject or retain H_m with $\boldsymbol{\mathcal{D}} = (\mathcal{D}_m; m = 1, 2, ..., M)$ is denoted by $\delta_m(\boldsymbol{\mathcal{D}}) \in \{0, 1\}$ or δ_m for short, where δ_m is 1 if H_m is rejected and is 0 otherwise. For ease of exposition, we denote the event that a null hypothesis is true (false) by $H_m = 0$ ($H_m = 1$). Table 1 contains our notation for the total number of rejected and retained null hypotheses, incorrectly rejected and retained null hypotheses, correctly rejected and retained null hypotheses, and number of true and false null hypotheses.

The objective of most multiple testing procedures is to define the decision functions $\delta = (\delta_m; m = 1, 2, ..., M)$ so as to maximize the expected number of true discoveries/positives $ETP = E[S]$, or minimize some type II error rate, such as the false non-discovery rate $FNR = E[U]/E[M - R]$ (Sun and Cai, 2007), subject to the constraint that the family-wise error rate $FWER = \Pr(V > 0)$ or false discovery rate $FDR = E[V/R|R > 0]\Pr(R > 0)$ is not more than a pre-specified level α. The FDR, or a variation of it such as the $mFDR = E[V]/E[R]$, is commonly utilized in large-scale multiple-hypothesis testing.

2.2 Weighted BH Methods

Many multiple testing procedures have been developed for p-value statistics $\boldsymbol{P} = (P_m; m = 1, 2, \ldots, M)$. The basic idea is to find a p-value threshold t for rejections and define $\delta_m(\boldsymbol{P}) = I(P_m \leq t)$ where $I(\cdot)$ is the indicator function. The well-known Benjamini and Hochberg (1995), or BH, procedure is implemented by finding the threshold $t_{BH} = \alpha k/M$ where $k = \max \left\{ m : P_{(m)} \leq \alpha m/M \right\}$ and $P_{(m)}$ is the mth order p-value. The BH procedure is then given by $\delta_m(\boldsymbol{P}) = I(P_m \leq t_{BH})$. Benjamini and Hochberg (1995) showed that if p-values from true null hypotheses are mutually independent and independent of p-values from false null hypotheses then this procedure has $FDR = \pi_0 \alpha \leq \alpha$, where $\pi_0 = M_0/M$ is the proportion of true null hypotheses. Adaptive FDR procedures (called ABH henceforth), leverage the fact that the BH procedure has FDR $= \pi_0 \alpha$ by estimating π_0 via $\hat{\pi}_0$ and apply the BH procedure at level $\alpha/\hat{\pi}_0$ (Storey et al., 2004). For example, Storey et al. (2004) proposed estimating π_0 and showed that if p-values are independent the ABH controls the FDR and is less conservative than the BH procedure.

Recent work has further improved upon the BH and ABH procedures by incorporating heterogeneity through p-value weighting. For example, letting $w_1, w_2, \ldots, w_M$ be weights satisfying $M^{-1} \sum_m w_m = 1$, the weighted BH procedure (WBH) in Roeder and Wasserman (2009) operates by applying the BH procedure to the weighted p-values denoted by $Q_m = P_m/w_m$. Roeder and Wasserman (2009) showed that the WBH procedure provides FDR control under a finite mixture model for the p-values considered in Genovese and

Table 1: Notation for various hypothesis testings subgroups based on if the null hypothesis is true ($H_m = 0$) or false ($H_m = 1$), and if the tests were rejected ($\delta_m = 1$) or not rejected ($\delta_m = 0$).

	$\delta_m = 0$	$\delta_m = 1$	Total
$H_m = 0$	T	V	M_0
$H_m = 1$	U	S	M_1
	$M - R$	R	M

Wasserman (2002), among others. Optimal weights for the WBH procedure can depend on heterogeneous prior probabilities for the states of the null hypotheses (Roeder and Wasserman, 2009; Hu et al., 2010; Li and Barber, 2017; Lei and Fithian, 2018; Zhang and Chen, 2020; Li and Barber, 2019), heterogeneous power functions (Peña et al., 2011) or both (Cai and Sun, 2009; Ignatiadis et al., 2016; Habiger, 2017; Ignatiadis and Huber, 2017).

Habiger (2017) proposed applying the adaptive BH procedure to weighted p-values. The procedure, henceforth called weighted adaptive BH (WABH), operates as follows: (i) compute weighted p-values via $Q_m = P_m/w_m$ with $\boldsymbol{Q} = (Q_m; m = 1, 2, \ldots, M)$, (ii) estimate $\hat{\pi}_0 = \{\sum_m I(Q_m \geq \kappa) + 1\}/\{M(1 - \kappa)\}$, (iii) compute threshold $t_{WABH} = \min\{\alpha, k\alpha/(\hat{\pi}_0 M)\}$ where $k = \max\{m : Q_{(m)} \leq m\alpha/(\hat{\pi}_0 M)\}$, and (iv) compute $\delta_m(\boldsymbol{Q}) = I(Q_m \leq t_{WABH})$. Habiger (2017) showed that for reasonably specified weights the WABH procedure controls the FDR asymptotically and has higher ETP than the WBH and ABH procedures. In particular, as long as the utilized weights are positively correlated with optimal weights the procedure still controls the FDR and is more powerful than unweighted procedures. This allows for a procedure that incorporates heterogeneity across tests in applications where the precise nature and degree of heterogeneity isn't well known, but may be estimated or reasonably specified. The first step in utilizing such a procedure is to specify optimal Oracle weights.

2.3 Optimal Oracle Weights

Optimal Oracle weights are allowed to depend on heterogeneous prior probabilities and/or power functions. Suppose, for example, P_m has null CDF $F_0(t) = \Pr(P_m \leq t | H_m = 0) = t$ and alternative CDF $F_m(t) = \Pr(P_m \leq t | H_m = 1)$. Further let $p_m = \Pr(H_m = 1)$ be the prior probability that H_m is non-null for $m = 1, 2, ..., M$. Suppose, for the moment, the Oracle situation where the weighted p-value threshold t along with p_m and F_m for each m are known. The weighted p-value decision rule can be written $\delta_m(Q_m) = I(Q_m \leq t) = I(P_m \leq w_m t) \equiv I(P_m \leq t_m)$.

The calculation of the optimal weights reduces to maximizing the expected number of

8

true positives, $ETP = \sum_m p_m F_m(t_m)$ subject to the constraint that $M^{-1} \sum_m t_m = t$. That is, the objective is to find

$$\max_{\{t_m : m = 1, \ldots, M\}} \left\{ \sum_m p_m F_m(t_m) \right\} \text{ such that } G(\boldsymbol{t}; \boldsymbol{p}, \alpha) = 0,$$

where $G(\boldsymbol{t}; \boldsymbol{p}, \alpha) = (1 - \alpha) \sum_m (1 - p_m) t_m - \alpha \sum_m p_m F_m(t_m)$ which can be solved via Lagrange multipliers (Habiger, 2017).

Assuming P_m arises from a normally distributed test statistic, the power of a test of size t is $F_m(t) = \bar{\Phi} \left\{ \bar{\Phi}^{-1}(t) - g_m \right\}$ where $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$, $\Phi(\cdot)$ is a standard normal distribution function, and g_m the *effect size over the standard error* of test m (defined in Section 3.2). The expression for $f_m(t) = \frac{d}{dt} F_m(t)$ is

$$f_m(t_m) = \frac{\phi\{\bar{\Phi}^{-1}(t_m) - g_m)\}}{\phi\{\bar{\Phi}^{-1}(t_m)\}}, \tag{1}$$

where $\phi(t)$ is the standard normal density function and

$$\frac{d}{dt_m} G(\boldsymbol{t}; \boldsymbol{p}, \alpha) = G'(t_m, p_m, \alpha) = (1 - p_m)(1 - \alpha) - \alpha p_m f_m(t_m).$$

Setting $p_m f_m(t_m) - \lambda G'(t_m, p_m, \alpha) = 0$ yields the following expression

$$f_m(t_m) = \frac{\lambda(1 - p_m)(1 - \alpha)}{p_m(1 + \lambda\alpha)} = c_m(\lambda),$$

where the solution in terms of t_m is

$$t_m(\lambda) = \bar{\Phi} \left[0.5 g_m + \log\{c_m(\lambda)\} g_m^{-1} \right]. \tag{2}$$

The Lagrange multiplier is found by solving

$$\sum_m (1 - p_m) t_m(\lambda) - \alpha \left[\sum_m (1 - p_m) t_m(\lambda) + \sum_m p_m F_m\{t_m(\lambda)\} \right] = 0 \tag{3}$$

for λ. The weights are then given by $w_m = t_m(\hat{\lambda})\{M^{-1} \sum_m t_m(\hat{\lambda})\}^{-1}$ where $\hat{\lambda}$ is the solution to (3). Once the weights are calculated the WABH procedure can be implemented.

2.4 Dependence between tests

Storey (2003) showed that the ABH procedure provides asymptotic FDR control under a weak dependence structure for the p-values and Habiger (2017) extended this result to weighted p-values. Weak dependence occurs, for example, when (weighted) p-values are correlated within groups but independent across groups. This structure may be reasonable to our application of interest because p-values are likely to be correlated within regions or clusters of voxels, but are nearly independent across distant regions. As a result, we ignore the dependence between p-values in our proposed testing procedure. Our simulation studies in Section 4 explore the impact of dependence on our proposed method and other procedures by generating spatially dependent locations for non-null tests (to varying degrees) with spatially dependent data. Those results suggest that the degree of dependence does not have an impact on the FDR of the proposed method.

3 Estimation of weights in VLSM

VLSM is a procedure that measures the strength of the association between lesion status and a cognitive outcome, independently for each voxel (Bates et al., 2003). Let X_{im} denote a measure of whether brain voxel m has a lesion for person i, $m = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, n$. Let Y_i denote the outcome of interest for person i, which we assume to be continuous. Further, let $X_{im}^+ = h(\sum_{j \neq m} X_{ij})$ be a measure of the total lesion size excluding voxel m for person i for some function h. Below we consider $X_{im}^+ \in \mathbb{R}$, but incorporating multidimensional X_{im}^+ is straightforward. Since voxel damage can only be a detriment to the cognitive outcome we consider H_m to be one-sided hypothesis tests, however, the methods are easily generalized to two-sided tests. We focus on logistic regression since it can model total lesion size (X_{im}^+) as a nuisance confounder (Karnath et al., 2004; Arnoux et al., 2018).

The optimal oracle weights in Section 2.3 require the specification of g_m and p_m in equations (1) and (3). To estimate the p_m's we use existing general methods (discussed in

Section 3.1). In Section 3.2, the heterogeneous g_m's are calculated by utilizing heterogeneous standard error calculations and prior knowledge that the power of the tests are low. Our resulting WABH algorithm is outlined in Section 2.

3.1 Estimation of prior null probabilities using known methods

Ideally, values for the prior null probabilities (p_m) can be based on previous studies and expert knowledge. When this is not possible, p_m can be estimated based on the unweighted p-values. While there are many approaches to estimating p_m in such cases, we focus on the AdaPT and CAMT procedures (Lei and Fithian, 2018; Zhang and Chen, 2020), due to the ease of implementing them in statistical software. Let $\boldsymbol{z}_m$ denote the so-called 'side-information' hypothesized to have an impact on p_m and/or f_m, the density of P_m under the alternative. Both AdaPT and CAMT use the two-groups model where $P_m | \boldsymbol{z}_m, H_m \sim (1 - H_m) f_0 + H_m f_m$ where f_0 is the density of P_m under the null and $p_m = \Pr(H_m = 1 | \boldsymbol{z}_m)$. To implement AdaPT or CAMT, we need to specify a parametric form for the relationship between (a) $\boldsymbol{z}_m$ and p_m, and (b) $\boldsymbol{z}_m$ and f_m. For both models, $\log\{p_m/(1 - p_m)\}$ is modeled via components of $\boldsymbol{z}_m$. For (b), in AdaPT we assume f_m is a beta density, and for CAMT we assume the ratio f_m/f_0 is a beta density. Specifically, a $\text{beta}(k_m, 1)$ where $\log(k_m)$ is modeled via components of $\boldsymbol{z}_m$. See Lei and Fithian (2018), and Zhang and Chen (2020) for details on their estimation procedures. After AdaPT or CAMT are implemented, the estimates of p_m are extracted and used in our testing procedure.

The WABH procedure can be implemented in various specifications of (a) and (b) above. For our simulation study, $\boldsymbol{z}_m$ consists of the 2×2 grid coordinates of test m, denoted by $\boldsymbol{z}_m^p$, and the predicted standard error (S_m) denoted by $\boldsymbol{z}_m^f$. In the simulation study for both AdaPT and CAMT, $\boldsymbol{z}_m^p$ and $\log\{p_m/(1 - p_m)\}$ are related using a linear combination of 5 degree natural cubic splines for each coordinate and their interaction. Further, $\log(k_m)$ was modeled using a 5 degree natural cubic spline on $\boldsymbol{z}_m^f$. See Section 5 for the specification of these relationships in our real data analysis.

11

3.2 Estimation of effect sizes and MMW criterion

In this section, we develop expressions for estimating g_m, which will be used to estimate optimal weights in subsequent sections. We consider a strictly continuous outcome $Y_i \in \mathbb{R}$ and a binary $X_{im} \in (0, 1)$ lesion indicator, which is modeled as a function of Y_i and X_{im}^+ with a logistic regression model.

Let $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, $\boldsymbol{X}_m^+ = \{X_{1m}^+, X_{2m}^+, \ldots, X_{nm}^+\}$ and $\boldsymbol{X}_m = \{X_{1m}, X_{2m}, \ldots, X_{nm}\}$ for $m = 1, \ldots, M$. We consider null hypotheses $H_m : \beta_{1m} = 0$ and alternative $H_m : \beta_{1m} > 0$ where

$$\text{logit}\{\Pr(X_{im} = 1 | Y_i, X_{im}^+)\} = \beta_{0m} + \beta_{1m} Y_i + \beta_{2m} X_{im}^+,$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. The Wald test p-values are $P_m = \bar{\Phi}(\hat{\beta}_{1m}/\widehat{SE}_m)$, where $\widehat{SE}_m$ is the estimated standard error of $\hat{\beta}_{1m}$. Clearly, the power of a test will depend upon β_{1m} and SE_m. Using previous results Væth and Skovlund (2004), the standard error of $\hat{\beta}_{1m}$ can be approximated via

$$SE_m = \left[\frac{(1 - R_m^2)}{n s_Y^2 \bar{X}_m (1 - \bar{X}_m)} \right]^{1/2} + o_p(n^{-1/2}) \tag{4}$$

where R_m^2 is the coefficient of determination for regressing $\boldsymbol{Y}$ on $\boldsymbol{X}_m^+$, $\bar{X}_m = \sum_i (X_{im})/n$, $s_Y^2 = \sum_i (Y_i - \bar{Y})^2/(n-1)$. Then,

$$\frac{\beta_{1m}}{SE_m} = \frac{\eta_m}{S_m} + o_p(n^{-1/2}) \tag{5}$$

where $\eta_m = \beta_{1m} s_Y$ and $S_m = [(1 - R_m^2)/\{n\bar{X}_m(1 - \bar{X}_m)\}]^{1/2}$. Assuming normality of the test statistics, given $g_m = \eta_m/S_m$ the power of a test of size t is $F_m(t) = \bar{\Phi}\left\{\bar{\Phi}^{-1}(t) - g_m\right\}$. Since S_m can be estimated *a priori*, the heterogeneity in the power can be calculated given the effect size η_m. To specify η_m, we consider the case where prior information differentiating η_m is unavailable and set $\eta_m = \eta$ for all m.

Figure 1 displays the relationship between the weights (w_m) and S_m for different η. This figure also shows the 10% rule (which is common in practice), where the weights are given by $w_m = MI(\bar{X}_m \in [0.1, 0.9])/(\sum_{k=1}^M I(\bar{X}_k \in [0.1, 0.9]))$. Note that the 10% rule up weights tests (i.e., $w_m \geq 1$) with high power (lower S_m) and down weight tests with low power.
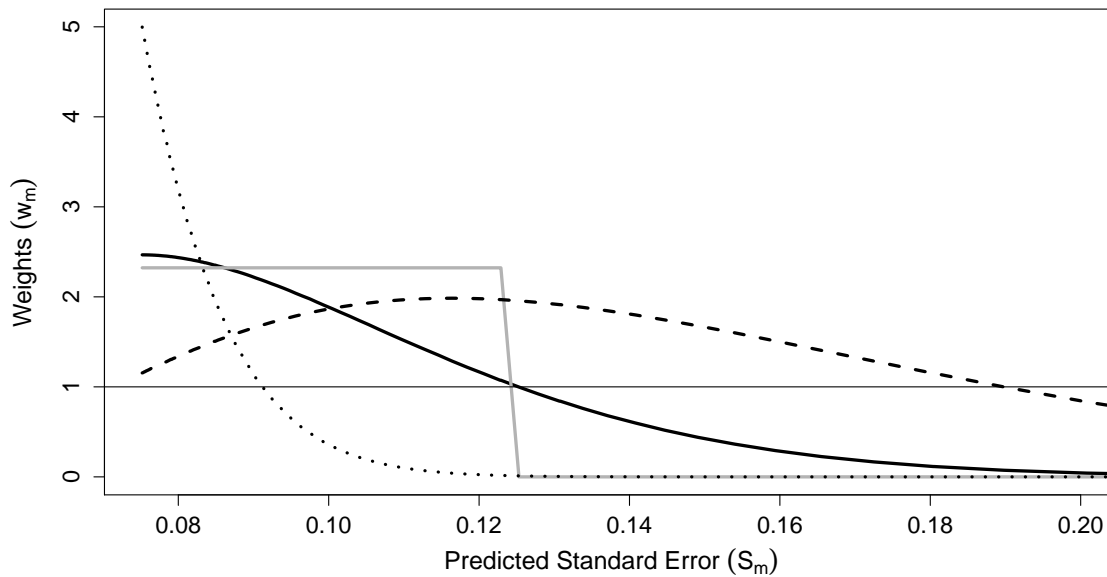
Figure 1: Estimated p-value weights (w_m) by their predicted standard error (S_m) for the aphasia data where $\eta = 0.1$ (black dotted), 0.178 (black solid), and 0.25 (black dashed), along with the 10% rule (gray solid).

That is, the weights are monotonically non-decreasing in power (or non-increasing in S_m). Habiger (2017) showed that in low-power settings, optimal weighting schemes do involve increasing weight for tests with higher power and decreasing weights for those with lower power. Thus, the intuition behind the 10% rule is correct, however, a more sophisticated weighting scheme from the optimal weights in Section 2.3 is available. We refer to our weights as the monotone minimum weights (MMW). In summary, the MMW weights are the specific optimal weights that satisfy the desired monotonicity property while ensuring that weights are not too aggressive, i.e. no weight is too large (aggressive weighting schemes may result in very large weights for only a few tests and amount to not testing the vast majority of tests, which is intractable).

Let us provide details. First, while S_m in equation (5) is a known source of heterogeneity affecting power, η_m is not. The MMW, for p_m's computed as in the previous section, arises by choosing η as large as possible so that weights are still monotone. The resulting weights

13

are depicted in Figure 1 by the thick black line. Note that in Figure 1 the MMW weights amount to choosing $\eta = .178$. Other values do not satisfy the MMW criteria. For example, choosing $\eta = 0.25$ results in weights that are not monotone. Choosing $\eta = 0.1$ results in monotone weights but use a smaller η than the MMW weights. Consequently, this results in a few large weights and most weights being 0 (i.e. non-robust weights). The general expression for MMW weights is provided in Theorem 3.1 below.

Theorem 1 *For a fixed λ and $p_m = \tau$ for all m, the maximum η such that $w_m - w_{m'} \geq 0$ for $S_m - S_{m'} \leq 0$ for all $m, m' \in \{1, \ldots, M\}$ is given by $\tilde{\eta} = S_{(1)} \sqrt{2 \log\{c(\lambda)\}}$ where $S_{(1)} = \min(S_m)$, $c(\lambda) = \lambda(1 - \tau)(1 - \alpha)/\{\tau(1 + \lambda\alpha)\}$, and $\log\{c(\lambda)\} > 0$. In this case, the thresholds are given by*

$$\tilde{t}_m(\lambda) = \bar{\Phi} \left\{ 0.5 \left(\frac{S_{(1)} \sqrt{2 \log\{c(\lambda)\}}}{S_m} \right) + \log(c_m) \left(\frac{S_{(1)} \sqrt{2 \log\{c(\lambda)\}}}{S_m} \right)^{-1} \right\}. \quad (6)$$

A proof of this theorem is given in Section A of the Supplemental Material.

Under the MMW criteria, the impact of weighting is minimized in that proportion of up-weighted tests is maximized among all η values with non-increasing weights for $S_m \geq S_{(1)}$ and $p_m = \tau$. Calculating the weights for the MMW criteria is as straightforward as the fixed η case. The MMW η is $\tilde{\eta} = S_{(1)}(2 \log\{c(\hat{\lambda})\})^{1/2}$ where $\hat{\lambda}$ is such that (3) holds when using (6) for the thresholds. The value of τ guarantees that the weights satisfy the MMW criteria for non-null probabilities equal to τ. This is a tuning parameter set by the investigator. As we demonstrate in our simulation studies, setting $\tau = \max(p_m)$ is effective for studies that are low in power. For studies with more robust power, we find setting τ to mean or qth percentile of the observed p_m's can give better results. It is important to note that weighted FDR methods control the FDR regardless of whether optimal weights are used and – as long as weights are reasonable – are more powerful than their un-weighted counterparts (Roeder and Wasserman, 2009; Habiger, 2017).

3.3 WABH algorithm

R code to run our proposed testing procedure along with data to replicate the results of our analysis in Section 5 are available on GitHub (McLain and Zheng, 2022). A general form for implementing the WABH procedure at level α is given in Algorithm 1. Our implementation uses logistic regression to obtain S_m and P_m in step 1. The functional form between $\boldsymbol{z}_m$ and p_m, and $\boldsymbol{z}_m$ and f_m is required for step 2. While our procedure only uses the estimates of p_m in step 2, specifying a plausible model for $\boldsymbol{z}_m$ and f_m is important since a poor model can have a negative impact on the estimates of p_m.

There are multiple measures of the impact of weighting such as the proportion of up-weighted tests (i.e., those with $w_m \geq 1$), the maximum weight, and the proportion of tests that are *inconclusive* (e.g., those with $w_m < 0.1$). Inconclusive tests are those that are essentially ignored by the testing procedure, i.e., they are likely to be not rejected due to their low weight. Reporting which tests are inconclusive is an important step in implementing the WABH (or other weighting procedures) so that the tests that were essentially not included in the testing procedure can be known.

4 Simulation Study

To test the properties of the proposed methods we performed simulation studies with logistic regression models. We considered a two-dimensional 100×100 grid of data. Let $\mathcal{S}_1$ index the set of false nulls. The coordinates of the tests in $\mathcal{S}_1$ were simulated from a zero-mean Gaussian random field (GRF) with $\Sigma_s(m, m') = \exp\{-(||\boldsymbol{z}_m^p - \boldsymbol{z}_{m'}^p||_2/s)^2\}$ where $\boldsymbol{z}_m^p$ is the two-dimensional coordinates for point m and $||\cdot||_2$ is the l_2-norm. The tests in $\mathcal{S}_1$ were those with the largest $K = ||\mathcal{S}_1||$ simulated values. The data were generated via

$$\text{logit}\{\Pr(X_{im} = 1|Y_i, b_i)\} = \alpha_0^* + \alpha_{0m}^* + \alpha_{1m}^* Y_i + b_i, \tag{7}$$

where $\alpha_0^* = -1$, α_{0m}^* follows a zero-mean GRF with covariance function $C^2 \Sigma_{50}(m, m')$, $\alpha_{1m}^* \sim U(0, 2\theta)$ for all $m \in \mathcal{S}_1$ ($\alpha_{1m}^* = 0$ otherwise), $b_i \sim N(0, 0.8^2)$, and $Y_i = 0.5b_i + \epsilon_i$ where $\epsilon_i \sim N(0, 0.5^2)$. Recall from (4), that the power heterogeneity is driven by $Var(X_{im}) =$

Algorithm 1 A general form for the WABH algorithm.
---
1. For $m = 1, 2, \ldots, M$

    (a) Compute $\bar{X}_m$ and R_m^2 to obtain S_m.

    (b) Compute the unweighted p-value P_m.

2. Implement AdaPT or CAMT on the unweighted p-values, extract estimates of p_m for all m.

3. Specify $\eta_m = \eta$ directly or specify τ and compute $\tilde{\eta} = S_{(1)}\sqrt{2\log\{c(\lambda)\}}$. Compute g_m.

4. Compute the optimal weights by

    (a) plugging (g_m, p_m) into (3) and solving for $\hat{\lambda}$, and

    (b) compute optimal weight $w_m = t_m(\hat{\lambda})\{M^{-1}\sum_m t_m(\hat{\lambda})\}^{-1}$ using $t_m(\hat{\lambda})$ in (2).

5. Compute weighted p-values $Q_m = P_m/w_m$.

6. Implement adaptive BH procedure:

    (a) find $k^* = \max\{m : Q_{(m)} \leq m\alpha/(\hat{\pi}_0 M)\}$ where $\hat{\pi}_0 = M^{-1}\sum_m(1 - p_m)$, and

    (b) compute $\delta_m(\boldsymbol{Q}) = I(Q_m \leq t_{WABH})$ where $t_{WABH} = \min\{\alpha, k^*\alpha/(\hat{\pi}_0 M)\}$.

$\bar{X}_m(1 - \bar{X}_m)$, which will be a function of α_{0m}^* in that extreme α_{0m}^* will have low $Var(X_{im})$. Thus, $C^2 = Var(\alpha_{0m}^*)$ controls the amount of power heterogeneity among the tests.

The simulation settings were varied over $C = 0.5$, 1.5, and 3, corresponding to low, moderate, and high power heterogeneity, respectively, $s = 0.01$, 5, and 10, corresponding to low, moderate, and high spatial clustering, respectively, $K = 100$ or 500, and the expected effect size $\theta = 0.25, 0.5$, or 0.75 for low, moderate, or high average power. All simulations used $n = 200$. The fitted model for the mth test was

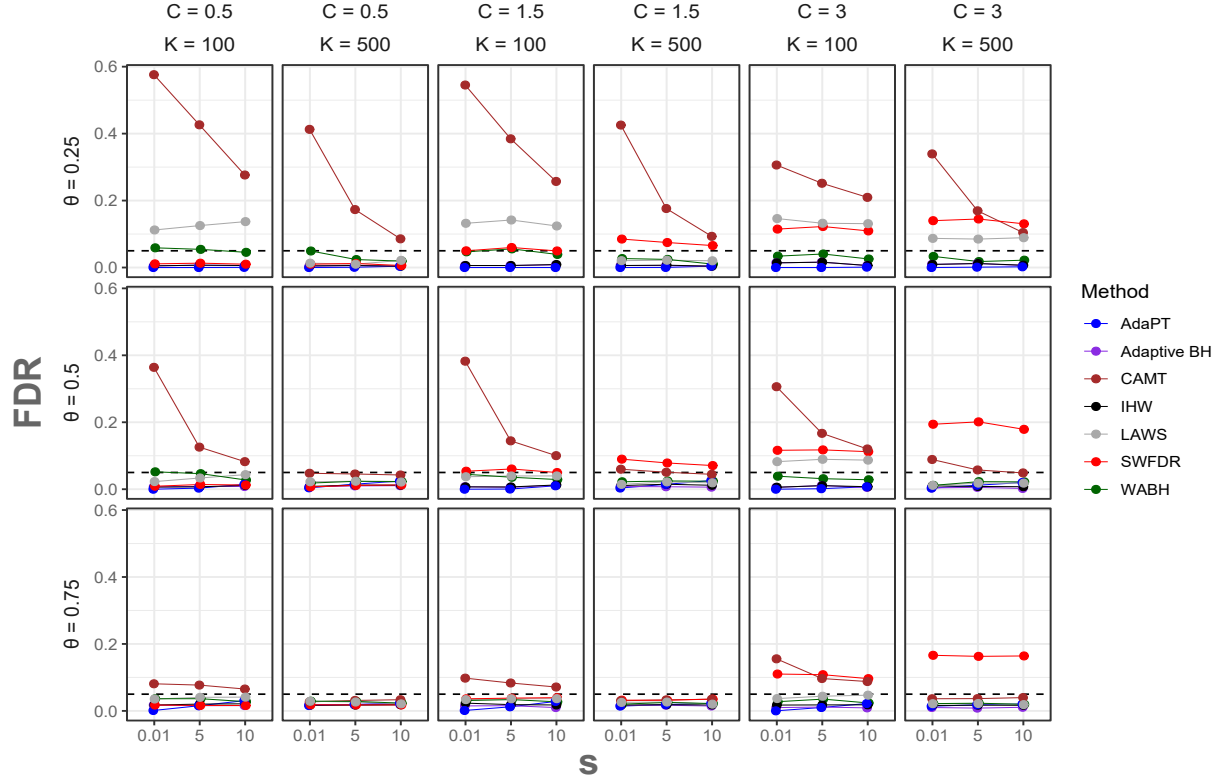$$\text{logit}\{\Pr(X_{im} = 1|Y_i, X_i^+)\} = \alpha_{0m} + \alpha_{1m}Y_i + \alpha_{2m}X_{im}^+$$

16

Figure 2: Average false discovery rates (FDR) by power heterogeneity (C), number of true signals (K), effect size (θ), and spatial dependence (s) for $M = 10000$ tests and $n = 200$.

where $X_{im}^{+} = \text{logit}\{(M-1)^{-1}\sum_{j\neq m} X_{ij}\}$. Section B of the Supplemental Material contains example plots of the data.

For all settings, the p-value weights were estimated using Algorithm 1 where p_m was estimated using AdaPT, or CAMT. For both AdaPT and CAMT,

$$\log\{p_m/(1-p_m)\} = \gamma_0 + \gamma_1 h_5(z_{m1}^p) + \gamma_2 h_5(z_{m2}^p) + \gamma_3 h_5(z_{m1}^p z_{m2}^p) \tag{8}$$

where z_{mj}^p is coordinate j of test m and $h_5(z)$ is a vector of 5 degree natural cubic splines with evenly spaced knots evaluated at z. Further, $\log(k_m) = h_5(S_m)$. To select η we used the MMW criteria with $\tau = 0.5$ or 0.9. We note that for AdaPT and CAMT the the beta assumption on the distribution of non-null p-values is likely misspecified. We also included the Adaptive BH, IHW, LAWS, and SWFDR procedures in the simulation. Adaptive BH was implemented with $\hat{\pi}_0$ being estimated using Storey (2007) with a threshold set at 0.05

17

as suggested for dependent data (Blanchard and Roquain, 2009). IHW was fitted with one covariate (S_m), as this was all the software allowed, with five-folds and automatic selection of the number of bins. SWFDR was fitted with a design matrix consisting of the two-dimensional coordinates and S_m to estimate the null probability for each test. To fit LAWS we used a threshold of 0.9 (the default) with a Gaussian kernel and bandwidth set to 4.5. LAWS does not model the non-null p-value distribution via covariates, thus S_m was not used.

The methods are compared in terms of $FDR = B^{-1} \sum_b I(R_b > 0)(V_b/R_b)$ and $Power = ETP/K$ ($ETP = B^{-1} \sum_b S_b$) where S_b, V_b and R_b denote the number of correct discoveries, false discoveries and total discoveries from the bth iteration. Procedures set FDR control level to $\alpha = 0.05$. All simulations were run for $B = 500$ iterations. For brevity, we show the WABH when p_m was estimated using CAMT with $\tau = 0.9$ only. Section B of the Supplemental Material contains results of WABH with AdaPT and other τ values along with other common methods.

In Figure 2, we present summarized FDR results of simulation studies. The AdaPT, Adaptive BH, IHW, and WABH procedures all controlled the FDR with values that are less than or close to the nominal level in all settings. The CAMT had high FDR when the proportion of false null hypotheses or expected effect size (θ) was low. For example, when $K = 100$ and $\theta = 0.25$, the FDR ranged from $0.2 - 0.6$ and $0.2 - 0.3$ in the low to moderate, and high heterogeneity settings, respectively. For $K = 100$ and $\theta = 0.75$, the CAMT FDR ranged from $0.06 - 0.16$, while for $K = 500$ the FDR was controlled reasonably well for $\theta \geq 0.5$. The LAWS procedure also resulted in FDR values that were above the nominal level when $\theta = 0.25$, and some $\theta = 0.5$ settings. Lastly, the FDR for the SWFDR procedure was above the nominal level in all high heterogeneity settings.

In Figure 3, we present summarized power results of simulation studies. Overall, among the procedures which control FDR in most settings (i.e., AdaPT, Adaptive BH, IHW, and WABH) the WABH procedure had the largest power. AdaPT had larger power for $\theta = 0.75$, $K = 500$ when $s \leq 5$. For $\theta = 0.75$, the LAWS procedure controlled the FDR and had
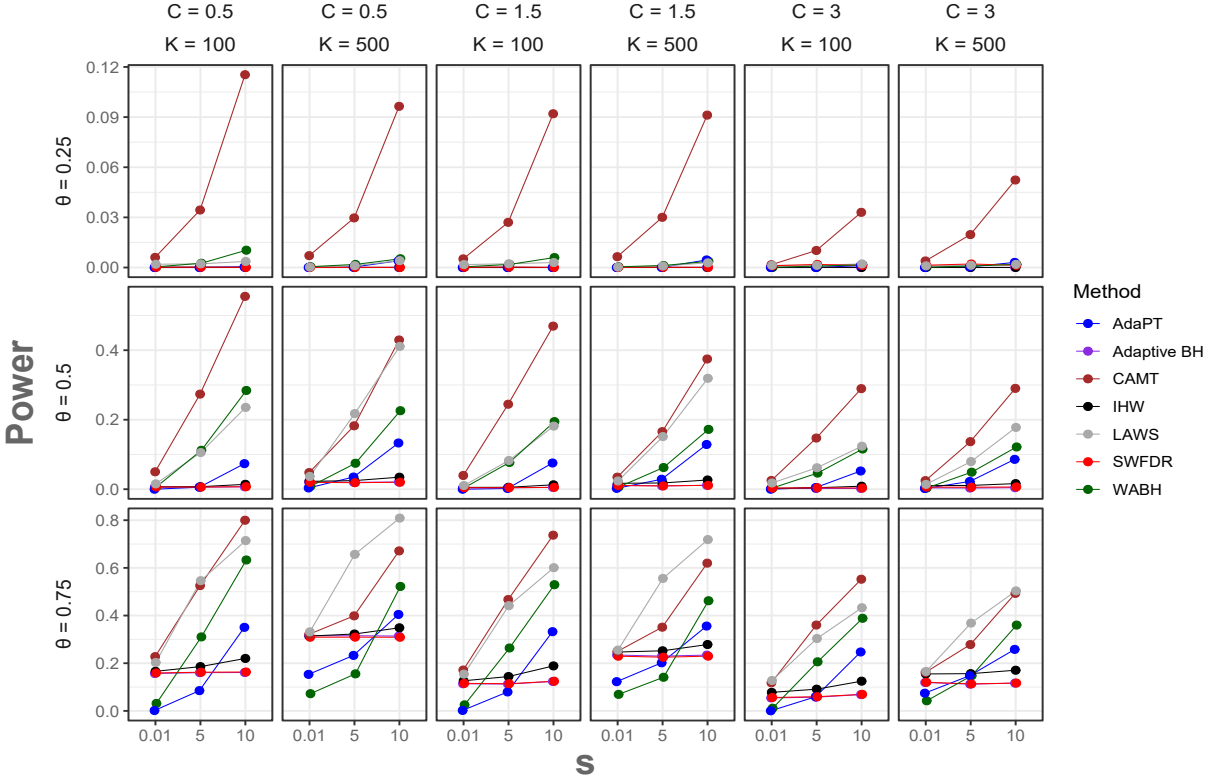
18

Figure 3: Estimated power (ETP/K) by the power heterogeneity (C), number of true signals (K), effect size (θ), and spatial dependence (s) for $M = 10000$ tests and $n = 200$.

the largest power (CAMT has larger power only when the FDR$\geq \alpha$). Further, LAWS performed well (FDR controlled and high power) for $\theta = 0.5$ and $K = 500$. As a result, the LAWS procedure appears to perform well in high-powered settings, however, such settings are not likely in our application of interest.

5 Data Analysis

Our sample was drawn from the population described in detail by Yourganov et al. (2015), and follows the same inclusion/exclusion criteria, behavioral testing, and behavioral analyses. In brief, all participants were adults with chronic left-hemisphere stroke and aphasia. All individuals were scanned using a 3T MRI scanner. Lesions were obtained by hand

by an expert on a high-resolution T2-scan for optimal identification of lesion boundaries. Lesions were coregistered to the individual's T1 scan. Each individual's lesion was warped to have a common size and shape through enantiomorphic normalization (Nachev et al., 2008) using our clinical toolbox (Rorden et al., 2012). Therefore, for each individual, the lesion was mapped as a binomial volume with a resolution of $181 \times 217 \times 181$ voxels (each 1mm^3), though many of these voxels are outside the brain or have zero damage for all subjects. For this study, we included the 220 individuals enrolled at the time of these analyses. Data from 834582 *candidate voxels* – voxels with damage to at least one subject – were considered.

To fit WABH we used CAMT to estimate p_m with the MMW criteria and $\tau = 0.9$. The side-information consists of $\boldsymbol{z}_m^p$ the 3-dimensional voxel coordinates and S_m. The relationship between $\boldsymbol{z}_m^p$ and p_m was similar to (8) with natural cubic splines on all three coordinates and all two-way interactions. We used 12 degrees of freedom for the splines which had the smallest BIC among the many values tested. The comparison methods included the 10% Rule, BH, ABH, AdaPT, and CAMT methods. For AdaPT, penalized regression splines with an automatic degree of freedom selection were used to relate $\boldsymbol{z}_m^p$ and p_m and S_m with f_m. We attempted an analysis with LAWS on this data also but it was not computationally feasible and failed to converge. Part of the issue is that LAWS requires a full cubic 3D structure. After removing slices with no candidate voxels more than 2×10^6 tests were still present (more than twice the tests of other procedures). For AdaPT and the WABH with AdaPT, we removed voxels with damage in less than 0.5% of subjects due to the convergence issue. Among voxels with non-zero damage, 360038 (43.1%) have damage in less than 5% of subjects. Codes and data to replicate the data analyses are available on GitHub McLain and Zheng (2022).

Lesion status was regressed as a function of the Aphasia Quotient (AQ) score (Y) and total lesion size (X^+). Checking the assumptions of all tests individually is not possible, so scatterplots of Y and X^+ were examined to determine the nature of the relationship between Y and X_m. This should suffice since X^+ is a measure of the average probability,
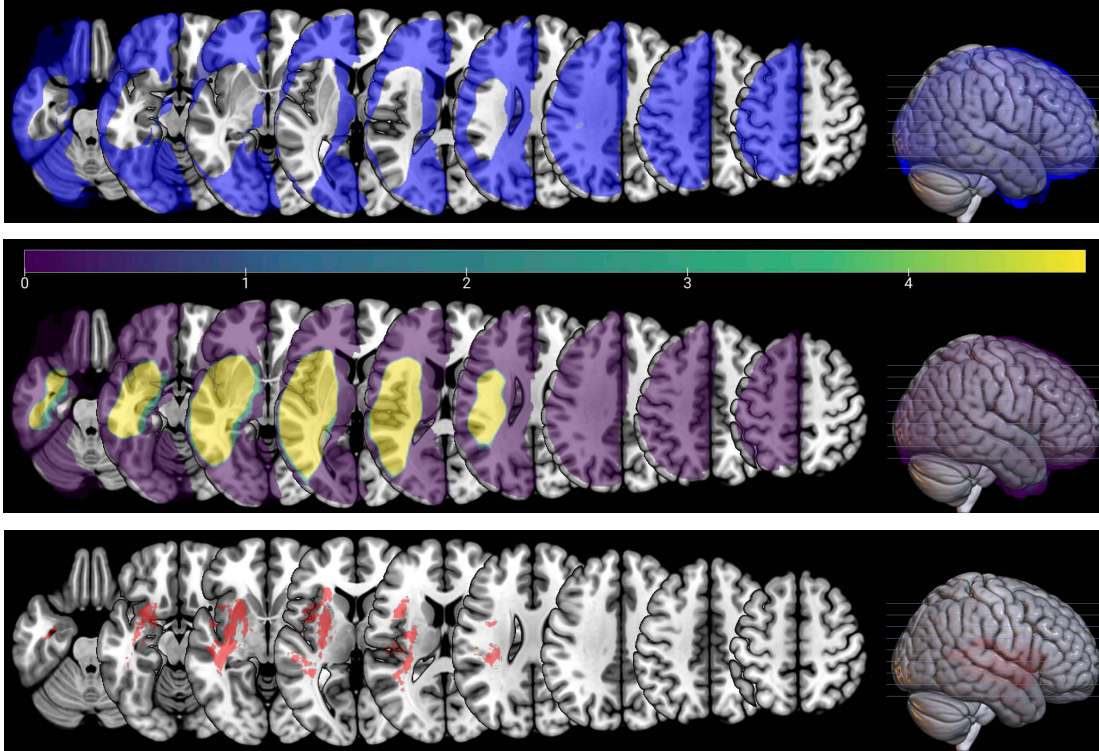
Figure 4: Inconclusive voxels (top, in blue), p-value weights (middle), and significant voxels (bottom, in red) for the WABH procedure. The plots are overlayed on a white structural brain image for reference.

and if the average probability is related to some transformation of Y then it is plausible that the voxel-level probabilities are related to the same transformation. Scatterplots of Y and X^+ showed a linear relationship when both were logit transformed. As a result, both Y and X^+ were included in the model after a logit transformation.

In Figure 4, we present a map of the estimated p-value weights, inconclusive voxels, and significant voxels for the WABH procedure with prior non-null probability estimated using CAMT. By knowing the weights, voxels, where the tests were inconclusive, can be shown. For example, the blue voxels in Figure 4 have weights less than 0.1 and thus could very likely contain type II errors. Such results are critical to show researchers which areas still require further study. Other weight metrics show that the WABH with AdaPT resulted in 36% up-weighted tests, 64% inconclusive tests, and 2.79 as the maximum weight. The

WABH procedure with CAMT has 23% up-weighted tests, 75% inconclusive tests, and 4.84 as the maximum weight.

The WABH with CAMT, AdaPT, and CAMT find 26568, 20, and 174953 significant voxels respectively, while WABH with AdaPT, 10% Rule, BH, and ABH find no significant voxels. Many of the significant voxels for WABH with CAMT appear to be located in and around the inferior frontal gyrus, which contains Broca's region which is a main area linked to speech production.

6   Discussion

In this paper, we proposed the use of weighted adaptive BH hypothesis testing for VLSM analysis. While the weighted adaptive BH procedure has been proposed by others, this paper was the first to incorporate heterogeneous prior non-null probability and proposed approach for estimating effect sizes in a manner consistent with anticipated low power assumptions of VLSM (see Theorem 3.1). The specific weighting scheme is available in Algorithm 1. Our simulation studies showed that our proposed method has a better performance than the other commonly used methods. Specifically, we found that while CAMT has high power the FDR was well above the nominal level, particularly in settings with a smaller expected effect size (θ) or a number of non-nulls (K). LAWS also had difficulty controlling the FDR for low values of θ. An in-depth assessment of why these methods – both of which have solid theoretical guarantees on their FDR values – fail to control the FDR is beyond the scope of this paper. However, it is evident that these methods have difficulty when θ and K are small. As a result, their inflated FDRs are in situations where estimating properties about p_m and f_m are challenging due to low power and/or a small number of non-null tests. The proposed method was the most powerful among those that controlled the FDR for most settings. The findings of the data analysis are consistent (though can't be confirmed) with the findings of the simulation study, where CAMT (AdaPT) found many more (less) significant voxels than our proposed method.

Our proposed WABH procedure ignored the dependence between the hypotheses tests.

The WABH procedure has been shown to provide asymptotic FDR control under a weak dependence structure on the p-values (Habiger, 2017), which is plausible for our setting. Benjamini and Yekutieli (2001) showed that the BH procedure still controls the FDR under a positive regression dependence structure (PRDS) and proposed modifications to the original BH procedure for arbitrary dependence. The PRDS property is satisfied if the test statistics are Gaussian, nonnegatively correlated and the testing hypotheses are one-sided. Since VLSM are usually one-sided tests and the spatial correlation between the test statistics will be (mostly) non-negative, the assumptions proposed by Benjamini and Yekutieli (2001) may be reasonable for the application of interest. However, extending the WABH to more general dependence scenarios is of interest.

In the data analysis, the number of discoveries was positively associated with the severity of weighting (i.e., heavier weighting resulted in more discoveries). However, heavy weighting results in many inconclusive hypotheses (up to 75% in our data analysis), and the regions are likely to include type II errors which need to be studied further. P-value weighting results in more discovered voxels by down-weighting voxels where discovery isn't likely and up-weighting voxels where it is. The result is more overall power in exchange for essentially not testing some voxels. It is important to acknowledge these later regions in reporting. This is why results such as Figure 4 should be included when they are employed so that the impact of weighting is transparent. Codes and data to replicate the data analyses are available on GitHub McLain and Zheng (2022).

Acknowledgements

References

Arnoux, A., Toba, M. N., Duering, M., Diouf, M., Daouk, J., Constans, J.-M., Puy, L., Barbay, M., and Godefroy, O. (2018), "Is VLSM a valid tool for determining the functional anatomy of the brain? Usefulness of additional Bayesian network analysis," *Neuropsychologia*, 121, 69 – 78.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018), "Weighted false discovery rate control in large-scale multiple testing," *Journal of the American Statistical Association*, 113, 1172–1183.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., and Dronkers, N. F. (2003), "Voxel-based lesion–symptom mapping," *Nature neuroscience*, 6, 448.

Benjamini, Y. and Cohen, R. (2017), "Weighted false discovery rate controlling procedures for clinical trials," *Biostatistics*, 18, 91–104.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 57, 289–300.

— (1997), "Multiple hypotheses testing with weights," *Scandinavian Journal of Statistics*, 24, 407–418.

Benjamini, Y. and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, 29, 1165–1188.

Blanchard, G. and Roquain, E. (2009), "Adaptive FDR control under independence and dependence," *J. Mach. Learn. Res.*, 10, 2837 –2831.

Boca, S. M. and Leek, J. T. (2018), "A direct approach to estimating false discovery rates conditional on covariates," *PeerJ*, 6, e6035.

Cai, T., Sun, W., and Xia, Y. (2021), "LAWS: A Locally Adaptive Weighting and Screening Approach to Spatial Multiple Testing," *Journal of the American Statistical Association*, 00, 1–14.

Cai, T. T. and Sun, W. (2009), "Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks," *J. Amer. Statist. Assoc.*, 104, 1467–1481.

Genovese, C., Roeder, K., and Wasserman, L. (2006), "False discovery control with p-value weighting," *Biometrika*, 93, 509–524.

Genovese, C. and Wasserman, L. (2002), "Operating characteristic and extensions of the false discovery rate procedure," *J. R. Stat. Soc. Ser. B Stat Methodol.*, 64, 499–517.

Habiger, J. D. (2017), "Adaptive false discovery rate control for heterogeneous data," *Statist. Sinica*, 27, 1731–1756.

Holmes, A. P., Blair, R., Watson, J., and Ford, I. (1996), "Nonparametric analysis of statistic images from functional mapping experiments," *Journal of Cerebral Blood Flow & Metabolism*, 16, 7–22.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010), "False discovery rate control with groups," *J. Amer. Statist. Assoc.*, 105, 1215–1227.

Ignatiadis, N. and Huber, W. (2017), "Covariate powered cross-weighted multiple testing with false discovery rate control," *arXiv preprint arXiv:1701.05179*, 1–58.

— (2021), "Covariate powered cross-weighted multiple testing," *Journal of the Royal Statistical Society Series B*, 83, 720–751.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016), "Data-driven hypothesis weighting increases detection power in genome-scale multiple testing," *Nature Methods*, 13, 577–580.

Karnath, H.-O., Fruhmann Berger, M., Küker, W., and Rorden, C. (2004), "The Anatomy of Spatial Neglect based on Voxelwise Statistical Analysis: A Study of 140 Patients," *Cerebral Cortex*, 14, 1164–1172.

Karnath, H.-O., Sperber, C., and Rorden, C. (2018), "Mapping human brain lesions and their functional consequences," *NeuroImage*, 165, 180 – 189.

Leek, J. T., Jager, L., Boca, S. M., and Konopka, T. (2021), *swfdr: Estimation of the science-wise false discovery rate and the false discovery rate conditional on covariates*, r package version 1.20.0.

Lei, L. and Fithian, W. (2018), "AdaPT: an interactive procedure for multiple testing with side information," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 649–679.

Li, A. and Barber, R. F. (2017), "Accumulation tests for FDR control in ordered hypothesis testing," *Journal of the American Statistical Association*, 112, 837–849.

— (2019), "Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 45–74.

McLain, A. C. and Zheng, S. (2022), "Weighted Adaptive BH Analyses," `https://github.com/alexmclain/WABH_Analysis`.

Nachev, P., Coulthard, E., Jäger, H. R., Kennard, C., and Husain, M. (2008), "Enantiomorphic normalization of focally lesioned brains," *Neuroimage*, 39, 1215–1226.

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

Peña, E. A., Habiger, J. D., and Wu, W. (2011), "Power-enhanced multiple decision functions controlling family-wise error and false discovery rates," *Ann. Statist.*, 39, 556–583.

Roeder, K. and Wasserman, L. (2009), "Genome-wide significance levels and weighted hypothesis testing," *Statist. Sci.*, 24, 398–413.

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., and Karnath, H.-O. (2012), "Age-specific CT and MRI templates for spatial normalization," *Neuroimage*, 61, 957–965.

Rorden, C. and Karnath, H.-O. (2004), "Using human brain lesions to infer function: a relic from a past era in the fMRI age?" *Nature Reviews Neuroscience*, 5, 812.

Rorden, C., Karnath, H.-O., and Bonilha, L. (2007), "Improving lesion-symptom mapping," *Journal of cognitive neuroscience*, 19, 1081–1088.

Storey, J. (2002), "A direct approach to false discovery rates," *J. R. Stat. Soc. Ser. B Stat Methodol.*, 64, 479 – 498.

— (2003), "The positive false discovery rate: a Bayesian interpretation and the q-Value," *Ann. Statist.*, 31, 2012 – 2035.

Storey, J. D. (2007), "The optimal discovery procedure: a new approach to simultaneous significance testing," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 69, 347–368.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 66, 187–205.

Sun, W. and Cai, T. T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," *J. Amer. Statist. Assoc.*, 102, 901–912.

Tukey, J. W. (1994), "The problem of multiple comparisons Volume VIII: Multiple comparisons, 1948-1983," in *The collected works of John W. Tukey*, ed. Braun, H. I., Chapman & Hall, pp. 1–300.

Væth, M. and Skovlund, E. (2004), "A simple approach to power and sample size calculations in logistic regression and Cox regression models," *Statistics in Medicine*, 23, 1781–1792.

Yourganov, G., Smith, K. G., Fridriksson, J., and Rorden, C. (2015), "Predicting aphasia type from brain damage measured with structural MRI," *Cortex*, 73, 203–215.

Zhang, X. and Chen, J. (2020), "Covariate adaptive false discovery rate control with applications to Omics-Wide multiple testing," *Journal of the American Statistical Association*, 00, 1–17.

Supplemental Material

A. Proof of Theorem 3.1

For a fixed $p_m = p$ for all m, to prove theorem 1 we wish to find the maximum η such that

$$w_m - w_{m'} \geq 0 \text{ for } S_m - S_{m'} \leq 0 \text{ for all } m, m' \in \{1, \ldots, M\}. \tag{9}$$

To show that $\tilde{\eta}$ in the text is the unique solution to (9), we first establish sufficient criteria for the weights to be non-increasing in S, then show that $\tilde{\eta}$ is the largest such η that satisfy this condition.

Let $t_m(\lambda) \equiv t(\lambda, \eta, S_m, p_m)$ where recall that

$$t(\lambda, \eta, S_m, p_m) = \bar{\Phi} \left[0.5 g_m + \log\{c(\lambda; p_m)\} g_m^{-1} \right]$$

where

$$c(\lambda; p_m) = \frac{\lambda(1 - p_m)(1 - \alpha)}{p_m(1 + \lambda \alpha)}$$

For a fixed $p_k = \tau$ for all k, note that $w_m - w_{m'} \geq 0 \Leftrightarrow t(\lambda, \eta, S_m, \tau) \geq t(\lambda, \eta, S_{m'}, \tau)$. Thus, we seek the largest η such that $t'(c, \eta, S_m, \tau) \leq 0$ for all $S_m \in [S_{(1)}, S_{(M)}]$, where $S_{(1)} = \min(S_m)$, $S_{(M)} = \max(S_m)$ and

$$
\begin{aligned}
t'(c, \eta, S_m, \tau) &= \left. \frac{d}{dS} t(c, \eta, S, \tau) \right|_{S = S_m} \\
&= -\phi \left[\left(\frac{\eta}{2 S_m} \right) + \log\{c(\lambda; \tau)\} \left(\frac{S_m}{\eta} \right) \right] \left[\frac{\log\{c(\lambda; \tau)\}}{\eta} - \frac{\eta}{2 S_m^2} \right]. \tag{10}
\end{aligned}
$$

Note that,

(i) if $\eta_m = S_m \sqrt{2 \log\{c(\lambda; \tau)\}}$ then $t'(c, \eta_m, S_m, \tau) = 0$.

Further, the derivative of the latter portion of (10) with respect to η is negative for all $\eta > 0$ (which is sufficient since the first portion is negative for all p, S and η), thus for $\eta_m = S_m \sqrt{2 \log\{c(\lambda; \tau)\}}$

(ii) $t'(\lambda, \eta_m - \epsilon, S_m, \tau) < 0$ for $0 < \epsilon < \eta_m$, and

29

(iii) $t'(\lambda, \eta_m + \epsilon, S_m, \tau) > 0$ for $\epsilon > 0$.

By (i) if $S_m - S_{m'} < 0$ then $\eta_m - \eta_{m'} < 0$ and by (ii) $t'(c, \eta_{m'}, S_m, \tau) < 0$. As a result, $t'(c, \tilde{\eta}, S_m, \tau) \leq 0$ for all $S_m \in [S_{(1)}, S_{(M)}]$ where $\tilde{\eta} = S_{(1)}\sqrt{2\log\{c(\lambda;\tau)\}}$. As a result, (9) holds for $\tilde{\eta}$. The fact that $\tilde{\eta}$ is the largest η to satisfy (9) follows from (iii).

B. Simulation Study Results

In Figure 5 and Figure 6, we present additional summarized results of simulation studies with more parameters settings for WABH (WABH procedures with CAMT estimated non-null probability and MMW η where $\tau = 0.5$ or 0.9, WABH procedure with AdaPT estimated non-null probability and MMW η where $\tau = 0.9$, WABH procedure with Storey constant non-null probability). All procedures have acceptable FDR values, and the WABH procedures with AdaPT or CAMT estimated non-null probability have FDR values which are near 0.05. The WABH procedure with constant non-null probability have similar conservative FDR values to the Regular BH, Adaptive BH and Ten Rule procedures. WABH procedures with CAMT estimated non-null probability have relatively larger power than the other WABH procedures. When the spatial dependence is large, the WABH procedure with AdaPT estimated non-null probability has larger power than Regular BH, but less power than the WABH with CAMT estimated non-null probability. The power of WABH procedure with constant non-null probability has not much difference with the Regular BH procedure. In Figure 7, we can find that the signals will change from randomly spread to clustered when the spatial dependence increases. Figure 8 shows the lesion status examples for low, moderate and high power heterogeneity.
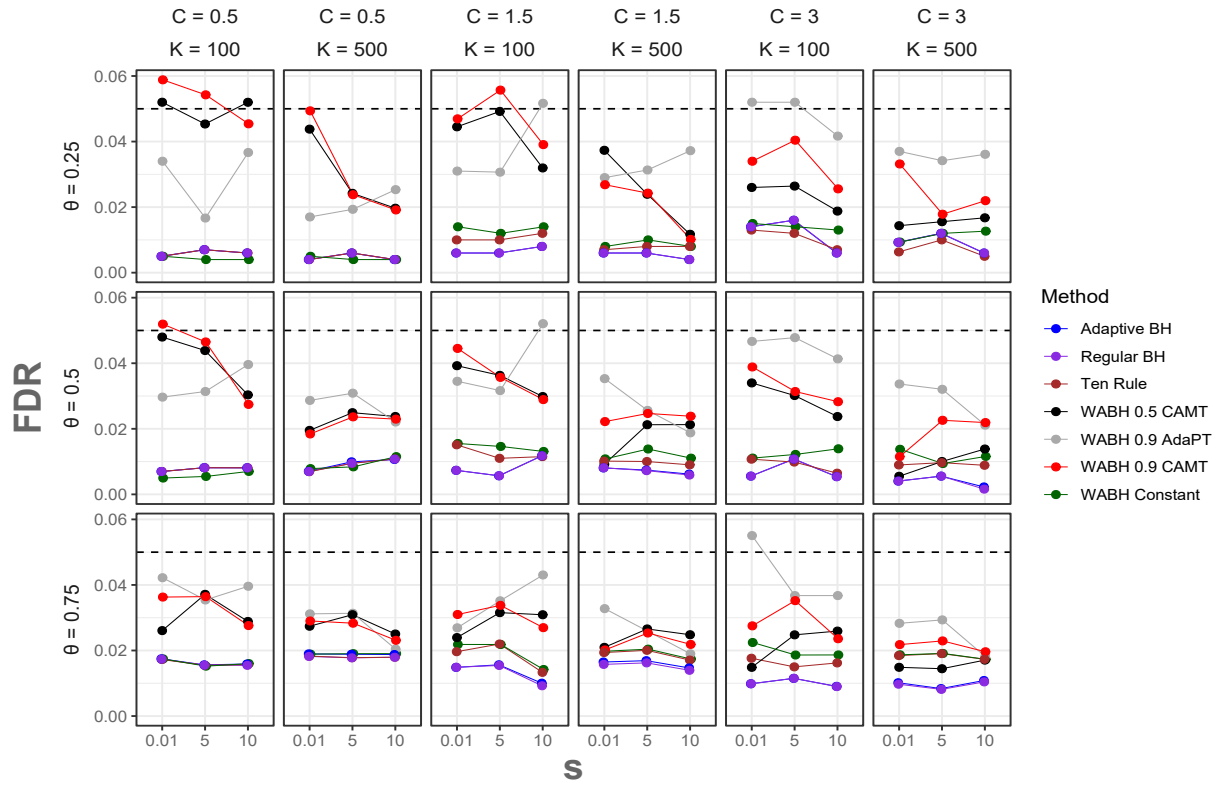
Figure 5: Average false discovery rates (FDR) by power heterogeneity (C), number of true signals (K), effect size (θ), and spatial dependence (s) for $M = 10000$ tests and $n = 200$.
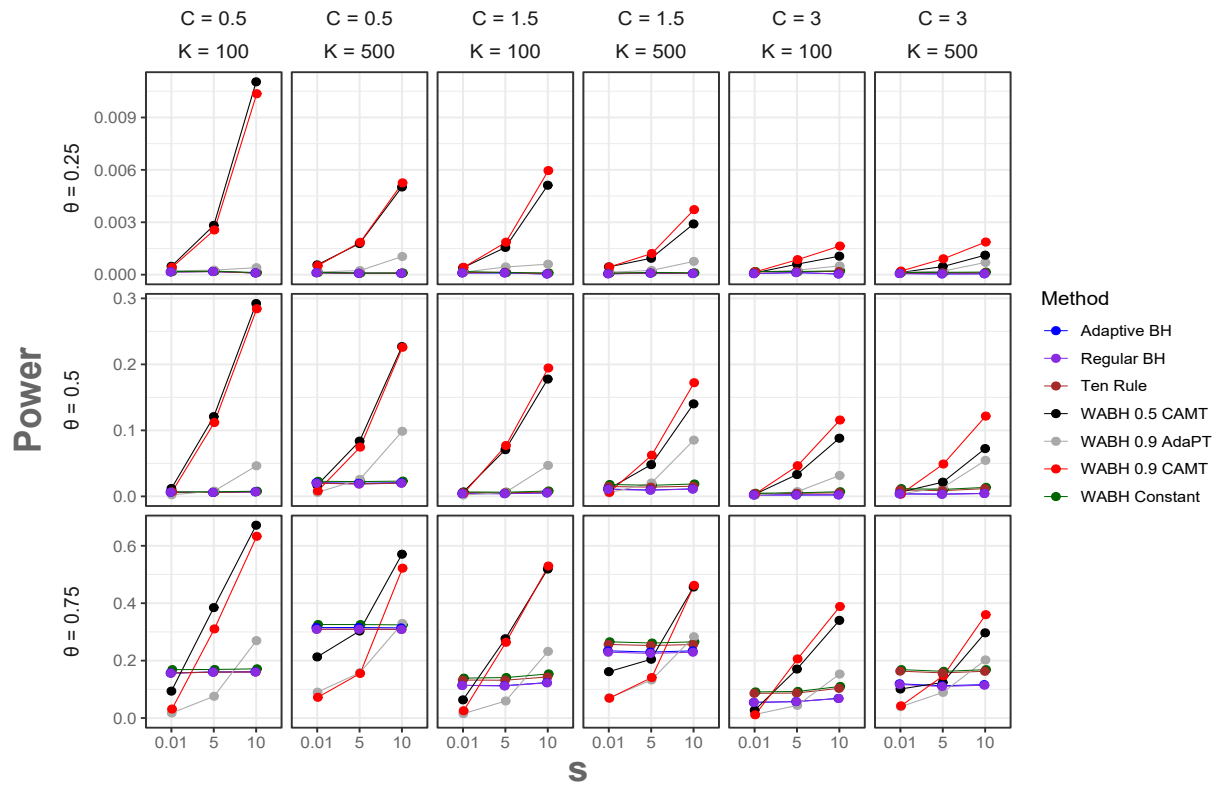
Figure 6: Estimated power (ETP/K) by the power heterogeneity (C), number of true signals (K), effect size (θ), and spatial dependence (s) for $M = 10000$ tests and $n = 200$.
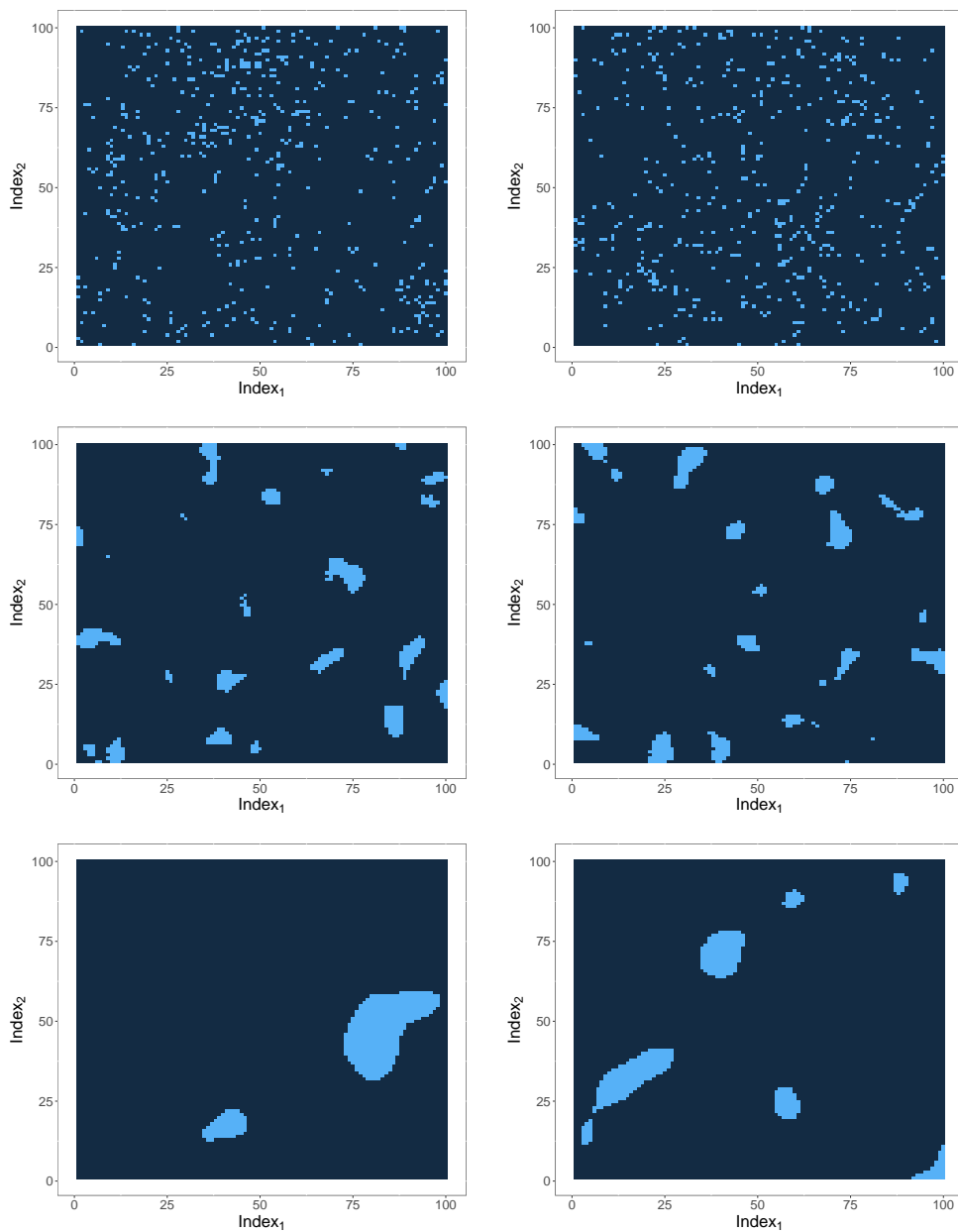
Figure 7: Examples of simulation signals (light blue) with low to high spatial dependence (top to bottom) for 500 true signals out of $M = 10000$ tests.
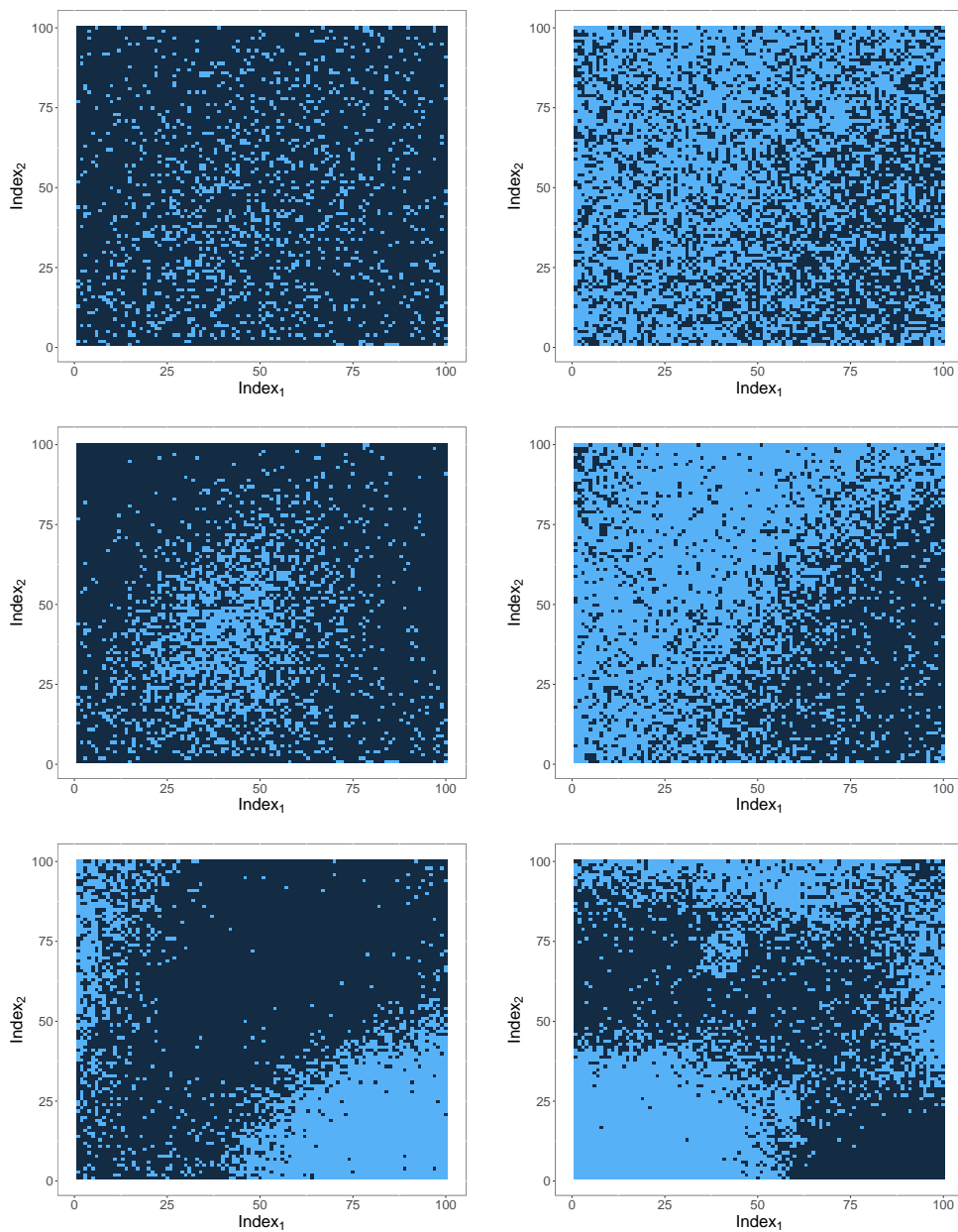
Figure 8: Examples of simulation lesion status with low to high power heterogeneity (top to bottom) for 500 true signals out of $M = 10000$ tests. Light blue refers to lesion.