NeuroImage 226 (2021) 117477

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Confidence Sets for Cohen's *d* effect size images

Alexander Bowring[a], Fabian J.E. Telschow[b], Armin Schwartzman[b,c], Thomas E. Nichols[a,d,e,*]

[a] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[b] Division of Biostatistics, University of California, San Diego, CA, USA
[c] Halicioğlu Data Science Institute, University of California, San Diego, CA, USA
[d] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[e] Department of Statistics, University of Warwick, Coventry, UK

ARTICLE INFO

Keywords:
Confidence sets
fMRI
Task fmri
Cohen's *d*
Effect sizes

ABSTRACT

Current statistical inference methods for task-fMRI suffer from two fundamental limitations. First, the focus is solely on detection of non-zero signal or signal change, a problem that is exacerbated for large scale studies (e.g. UK Biobank, $N = 40,000$+) where the 'null hypothesis fallacy' causes even trivial effects to be determined as significant. Second, for any sample size, widely used cluster inference methods only indicate regions where a null hypothesis can be rejected, without providing any notion of spatial uncertainty about the activation. In this work, we address these issues by developing spatial Confidence Sets (CSs) on clusters found in thresholded Cohen's *d* effect size images. We produce an upper and lower CS to make confidence statements about brain regions where Cohen's *d* effect sizes have exceeded and fallen short of a *non-zero* threshold, respectively. The CSs convey information about the magnitude and reliability of effect sizes that is usually given separately in a *t*-statistic and effect estimate map. We expand the theory developed in our previous work on CSs for %BOLD change effect maps (Bowring et al., 2019) using recent results from the bootstrapping literature. By assessing the empirical coverage with 2D and 3D Monte Carlo simulations resembling fMRI data, we find our method is accurate in sample sizes as low as $N = 60$. We compute Cohen's *d* CSs for the Human Connectome Project working memory task-fMRI data, illustrating the brain regions with a reliable Cohen's *d* response for a given threshold. By comparing the CSs with results obtained from a traditional statistical voxelwise inference, we highlight the improvement in activation localization that can be gained with the Confidence Sets.

1. Introduction

Online dating has transformed the love-seeking game forever. Whereas romantic partners would historically first encounter each other face-to-face, brought together by a mutual friend or family member, in recent times these rituals of connection have been largely replaced by social networks and matchmaking websites (Rosenfeld et al., 2019). While it was perhaps inevitable that technologies of the Digital Age would take a hold on our pursuit to find a partner, what may be more surprising is the influence the internet has had on the final outcomes of a marriage itself. In an investigation analyzing survey data from over 19,000 married American respondents, it was reported that virtual dating avenues may have helped to improve the prospects of finding a long and happy relationship (Cacioppo et al., 2013). With overwhelming statistical evidence, the results of this study found that spouses who had met their partner online were more likely to be satisfied with their marriage ($p < 0.001$) and less likely to divorce ($p < 0.002$).

Under closer inspection, however, these results are not all that they may seem. After this research was published, it was pointed out that the actual sizes of the observed effects were tiny (Nuzzo, 2013; 2014). Specifically, although the data had shown higher levels of marriage happiness for couples who met online compared to offline, the difference in means was from 5.5 to 5.6 on a 7-point scale; in terms of divorce rates, the deviation between groups worked out as one more break-up for every 100 marriages.

In this case study, the outcomes of the experiment were misconstrued due to an infamous pitfall with statistical testing known as the 'fallacy of the null hypothesis' (Rozeboom, 1960). The problem stems from the fact that statistical models conventionally assume mean-zero noise and that, under the null hypothesis, no signal is present. In practice, these assumptions are never completely upheld: all sources of noise will never completely cancel, and there will always be some (non-zero) signal everywhere. Consequentially, the smallest of effects will *always* become statistically significant given a sufficiently large sample size (Meehl, 1967),

* Corresponding author at: Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK.
E-mail address: thomas.nichols@bdi.ox.ac.uk (T.E. Nichols).

https://doi.org/10.1016/j.neuroimage.2020.117477
Received 22 May 2020; Received in revised form 13 October 2020; Accepted 16 October 2020
Available online 6 November 2020

even if they have little practical value or are unlikely to be replicable across repeated analyses (Button et al., 2013).

This issue has become topical within the functional magnetic resonance imaging (fMRI) community due to the arrival of population-scale neuroimaging datasets. While fMRI has traditionally been a 'small N' enterprise, with typical sample sizes of 20 to 30 subjects (Poldrack et al., 2017), datasets such as the Human Connectome Project (HCP, $N = 1200$) and UK Biobank ($N = 40,000+$) are now giving researchers the opportunity to analyze data acquired from tens of thousands of participants. These projects promise to transform our understanding of brain function, and are already yielding rich results (Miller et al., 2016; Van Essen and Glasser, 2016). However, in this setting the standard stastical thresholding approach to functional brain data analysis has become obsolete: with ample power to detect all effects, statistical analysis of high quality fMRI data has been shown to lead to almost universal brain activation, even when stringent thresholding methods are applied (Gonzalez-Castillo et al., 2012).

In a more general analysis context, there are still a number of limitations with traditional fMRI inference techniques. Currently, the most popular method for overcoming the multiple comparisons problem is to threshold the statistical results with cluster-extent based thresholding (Carp, 2012), involving a two-step procedure: first a primary voxelwise threshold is applied to the statistic map, usually in correspondence with an uncorrected significance level (e.g. $\alpha = 0.001$), creating clusters of voxels whose statistic values have all surpassed the threshold. Then, in order to control the family-wise error (FWE) rate, a cluster-extent threshold k is determined based on the distribution of cluster sizes obtained under the null-hypothesis of no activation, and the final results are computed as all suprathreshold clusters with a spatial extent larger than k.

As the FWE-corrected p-value is determined by cluster size, one of the main drawbacks of this procedure is that the significance of specific voxels can not be determined, and the most we can assert is that activation has occured *somewhere* inside a given cluster (Woo et al., 2014). It is therefore impossible to pinpoint the precise source of the activation when a cluster covers multiple anatomical regions, and the spatial specificity of the inference diminishes the larger a cluster becomes. Another problem with this approach is that no information is provided in regards to the spatial variation of significant clusters. For instance, if a single fMRI study was repeated many times using different groups of participants, there would be variation in the sizes and shapes of the final activation clusters, yet current statistical results have no way to convey this variability.

In our previous work (Bowring et al., 2019, (*BTSN*)), we helped to address these issues by developing Confidence Sets (CSs) for inference on %BOLD change effect size maps. Unlike traditional hypothesis testing methods, where inference is only provided in terms of the presence of an effect, the CSs made simultaneous confidence statements about the precise brain regions where raw effect sizes had exceeded, and fallen short of, a *non-zero* %BOLD threshold. Here, we set out to adapt the CSs for application to standardized Cohen's d maps (i.e. %BOLD change divided by population standard deviation) that are more commonly used to provide effect size estimates complementing the statistical results obtained from an fMRI one-sample t-test. For a cluster-forming threshold c and a predetermined confidence level $1 - \alpha$, the Cohen's d CSs comprise of two sets: the upper CS (denoted $\hat{\mathcal{A}}_c^+$), containing all voxels declared to have a true Cohen's d effect size greater than c; and the lower CS ($\hat{\mathcal{A}}_c^-$), for which all voxels *outside* this set are declared to have a true Cohen's d effect size less than c. The upper CS is smaller and nested inside the lower CS, and the assertion is made with $(1 - \alpha)100\%$ confidence holding simultaneously for both regions. Fig. 1 provides a visual schematic of the CSs, with the upper and lower CSs presented in red and blue respectively. Note that t-tests, or any test statistics, are *not* suitable for building CSs, as they do not estimate a population quantity and become arbitrarily large for increasing sample sizes.

The statistical characteristics of Cohen's d effect size maps are fundamentally different to the raw %BOLD images that motivated the CSs in our previous work. Our main contributions with this effort are modifications to the methods used in *BTSN* to create procedures for obtaining Cohen's d CSs on fMRI data with desirable finite-sample performance. In particular, we apply recent results from the bootstrapping literature and a variance-stabilizing transformation method to ultimately propose three separate algorithms for computing Cohen's d CSs. The first algorithm is motivated by asymptotic properties of the Cohen's d sampling distribution, and provides a framework for the two remaining methods which employ further adjustments to optimize the finite-sample performance of the CSs. We assess the performance of all three methods on a range of simulated synthetic 2D and 3D signals representative of fMRI clusters, and find that the two latter procedures are effective even when the sample size is modest ($N = 60$). Finally, we apply the three procedures to Human Connectome Project working memory task data, operating on Cohen's d effect maps, where we obtain CSs for a variety of cluster forming thresholds. By comparing the CSs with results obtained from a traditional statistical voxelwise inference, we highlight the improvement in activation localization that can be provided with the Confidence Sets.

The remainder of this manuscript is organized as follows: first, we describe the problem of obtaining Confidence Sets for Cohen's d images, exemplifying the key differences which distinguish Cohen's d from the %BOLD effect size. We then derive properties of the Cohen's d estimator, before adapting the methods developed in *BTSN* to propose three separate algorithms to compute Cohen's d CSs. We assess the empirical coverage performance of each of these methods on 2D and 3D Monte Carlo simulations, and finally, present the CSs obtained from applying each algorithm to Human Connectome Project working memory task-fMRI dataset.

2. Theory

2.1. From % BOLD to Cohen's d

For a compact domain $S \subset \mathbb{R}^D$, e.g. $D = 3$, for $i = 1, \dots, N$ consider the one-sample model at location $\boldsymbol{s} \in S$,

$$Y_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon_i(\boldsymbol{s}), \tag{1}$$

where $Y_1(\boldsymbol{s}), \dots, Y_N(\boldsymbol{s})$ are the observations at $\boldsymbol{s}$, $\mu(\boldsymbol{s})$ is the true underlying mean intensity across the observations, and $\epsilon_1(\boldsymbol{s}), \dots, \epsilon_N(\boldsymbol{s})$ are i.i.d. mean-zero errors with common variance $\sigma^2(\boldsymbol{s})$ and some unspecified spatial correlation. We are motivated by the setting of a group-level task-fMRI analysis, where $\mu(\boldsymbol{s})$ represents the true mean %BOLD change across the group, and each observation $Y_i(\boldsymbol{s})$ is the %BOLD response estimate map obtained by applying a first-level model to the ith participant's functional data. (Note, while we focus on the one-sample model here, the method may also generalize for application to the general linear model $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s})$. See the end of Section 5.1 for more details.)

We wish to make inference on the Cohen's d effect size, defined as the true mean %BOLD change divided by the population standard deviation,

$$d(\boldsymbol{s}) = \frac{\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}. \tag{2}$$

Specifically, we are interested in the brain regions where $d(\boldsymbol{s})$ has exceeded, and fallen short of, a fixed threshold c, indicated by the noise-free, population cluster defined as:

$$\mathcal{A}_c = \{\boldsymbol{s} \in S : d(\boldsymbol{s}) \geq c\}. \tag{3}$$

Since $\mathcal{A}_c$ is unknown, we pursue a method for constructing pairs of spatial CSs: an upper set $\hat{\mathcal{A}}_c^+$ and a lower set $\hat{\mathcal{A}}_c^-$, that we are confident surround the true excursion set $\mathcal{A}_c$ (i.e. $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$) for a desired confidence level of, for example, $1 - \alpha = 95\%$. Such a method lets us assert with 95% confidence that all voxels *contained* in the upper CS $\hat{\mathcal{A}}_c^+$

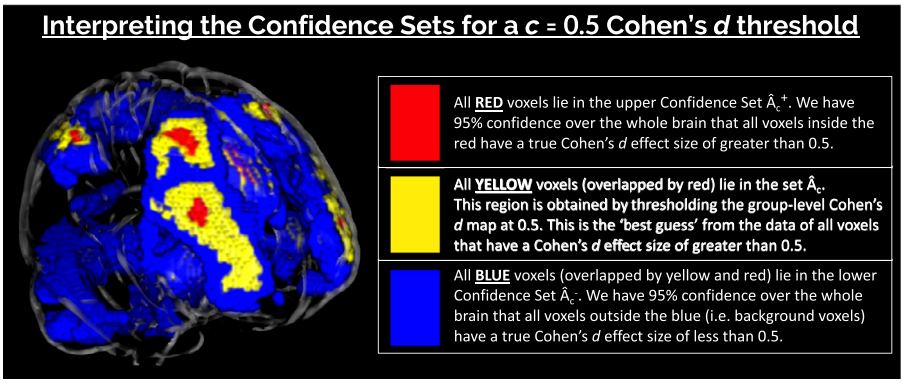


Fig. 1. Schematic of the color-coded regions we will use to visually represent the Confidence Sets (CSs) and point estimate set. The upper and lower CSs are presented in red and blue (overlapped by yellow and red) respectively. The yellow set (overlapped by red), $\hat{\mathcal{A}}_c$, is the point estimate set, the best guess from the data of voxels that have a Cohen's d effect size greater than the threshold $c = 0.5$.

have a Cohen's d effect size *greater* than, for example, $c = 0.8$, and simultaneously, we are 95% confident all voxels *outside* the lower CS $\hat{\mathcal{A}}_c^-$ have a Cohen's d effect size *less* than 0.8. Here, we emphasize the classic frequentist connotation of the term 'confidence'; letting $\partial\mathcal{A}_c$ denote the boundary of $\mathcal{A}_c$, then precisely, there is a probability of $1-\alpha$ that the region $\hat{\mathcal{A}}_c^- \cap (\hat{\mathcal{A}}_c^+)^c$ computed from a future experiment fully encompasses the true set boundary $\partial\mathcal{A}_c$. In this sense, the set difference of the upper and lower CS, $\hat{\mathcal{A}}_c^- \cap (\hat{\mathcal{A}}_c^+)^c$, is similar to a standard confidence interval.

In *BTSN*, we adapted the mathematical theory first proposed in Sommerfeld et al. (2018) (*SSS*) to obtain CSs for inference on the mean %BOLD change effect $\mu(s)$. Let $\bar{Y}(s) = \frac{1}{N}\sum_{i=1}^N Y_i(s)$, the sample mean %BOLD change. Then subject to continuity of the relevant fields and some basic conditions on the error terms $\epsilon_i(s)$, for the excursion set $\mathcal{A}_{c,\mu}$ of voxels with a true %BOLD effect size greater than c,

$$\mathcal{A}_{c,\mu} = \{s \in S : \mu(s) \geq c\}, \tag{4}$$

we showed that for a critical constant k, the upper and lower CSs constructed as

$$\hat{\mathcal{A}}^+_{c,\mu} := \left\{\mathbf{s} : \overline{Y}(\mathbf{s}) \geq c + \frac{k}{\sqrt{N}}\hat{\sigma}(\mathbf{s})\right\}, \hat{\mathcal{A}}^-_{c,\mu} := \left\{\mathbf{s} : \overline{Y}(\mathbf{s}) \geq c - \frac{k}{\sqrt{N}}\hat{\sigma}(\mathbf{s})\right\} \tag{5}$$

give asymptotic nominal coverage for enveloping the true $\mathcal{A}_{c,\mu}$ in terms of the mean %BOLD change effect size. Further to this, we proposed a Wild t-Bootstrap method for determining the critical value k, and demonstrated that on applying this method the CSs were also valid for data with smaller sample sizes.

We now seek to develop a similar methodology for the Cohen's d effect size. However, the statistical properties of the Cohen's d estimator $\hat{d}(s) = \frac{\bar{Y}(s)}{\hat{\sigma}(s)}$ are considerably different to the sample mean $\bar{Y}(s)$. To provide a visual intuition of this in the case of Gaussian data, in Fig. 2 we display images of both of these fields from a 2D simulation over a square region $S = 100 \times 100$. For $N = 60$ subjects, we simulated a toy run of the signal-plus-noise model in (1) where the true underlying signal $\mu(s)$ was a linear ramp effect increasing from a magnitude of 0 to 10 in the x-direction while remaining constant in the y-direction (Fig. 2(a)). To the signal we added subject-specific Gaussian noise $\epsilon_i(s)$, smoothed with a 3 voxel FWHM Gaussian kernel and then re-normalized to have a spatially constant standard deviation of $\sigma(s) = 1$. Notably, in this setup the true Cohen's d field $d(s)$ was identical to $\mu(s)$.

In Fig. 2(b) and (c) we show the sample mean and sample Cohen's d fields from this simulation. While the sample mean image is uniformly smooth across the space, the Cohen's d field becomes more speckled, i.e. more variable, for increasing true $d(s)$. In the following sections we will show that the sample Cohen's d is a biased estimator of the true underlying effect size, and that the sample variance of Cohen's d changes systematically with $d(s)$, before proposing our theoretical adjustments to the methods presented in *BTSN* to obtain CSs for Cohen's d effect size images.

2.2. Limiting properties of the Cohen's d estimator

Motivated by the example in Fig. 2, we now consider the one-sample model given in (1) with the additional assumption that the error fields are Gaussian. In this case, the data are i.i.d. $Y_1(s), \ldots, Y_N(s) \sim \mathcal{N}(\mu(s), \sigma^2(s))$ and the error terms $\epsilon_1(s), \ldots, \epsilon_N(s)$ are i.i.d. from a mean zero Gaussian random field $\epsilon(s)$ such that for all $s, t \in S$,

$$\mathrm{Cov}[\epsilon(s), \epsilon(t)] = \sigma(s)\sigma(t)\rho(s,t), \tag{6}$$

where $\rho(s,t)$ denotes the population correlation coefficient between points s and t in the error field. The sample mean and sample variance for this model are defined as

$$\bar{Y}(s) = \frac{1}{N}\sum_{i=1}^N Y_i(s), \quad \hat{\sigma}^2(s) = \frac{1}{N}\sum_{i=1}^N \Big(Y_i(s) - \bar{Y}(s)\Big)^2, \tag{7}$$

respectively.

We wish to understand the limiting structure of the Cohen's d estimator $\hat{d}(s) = \frac{\bar{Y}(s)}{\hat{\sigma}(s)}$. Applying the multivariate central limit theorem to the sample moments at s and t yields the asymptotic joint distribution:

$$\sqrt{N}\Big(\Big(\bar{Y}(s), \hat{\sigma}^2(s), \bar{Y}(t), \hat{\sigma}^2(t)\Big) - \Big(\mu(s), \sigma^2(s), \mu(t), \sigma^2(t)\Big)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \Sigma(s,t)\Big). \tag{8}$$

where the covariance matrix $\Sigma(s,t)$ is given by

$$\Sigma(s,t) = \begin{pmatrix} \sigma^2(s) & 0 & \sigma(s)\sigma(t)\rho(s,t) & 0 \\ 0 & 2\sigma^4(s) & 0 & 2\sigma^2(s)\sigma(t)^2\rho^2(s,t) \\ \sigma(s)\sigma(t)\rho(s,t) & 0 & \sigma^2(t) & 0 \\ 0 & 2\sigma^2(s)\sigma^2(t)\rho^2(s,t) & 0 & 2\sigma^4(t) \end{pmatrix}. \tag{9}$$

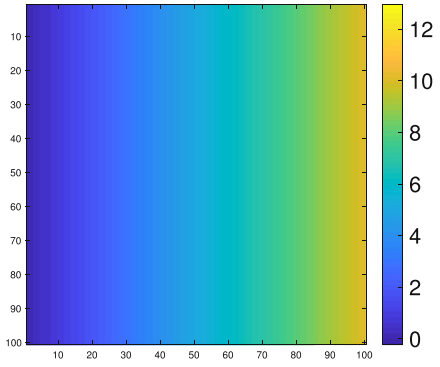
For the function $g(x_1, y_1, x_2, y_2) = \left(\frac{x_1}{\sqrt{y_1}}, \frac{x_2}{\sqrt{y_2}}\right)$, application of the delta method yields

$$\sqrt{N}\Big(\Big(\hat{d}(s), \hat{d}(t)\Big) - \Big(d(s), d(t)\Big)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \Sigma^*(s,t)\Big), \tag{10}$$

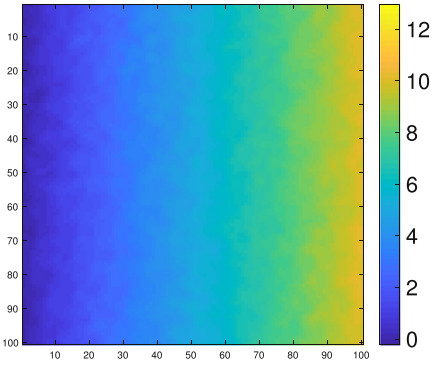
where

$$\Sigma^*(s,t) = \begin{pmatrix} 1 + \frac{d^2(s)}{2} & \rho(s,t) + \rho^2(s,t)\frac{d(s)d(t)}{2} \\ \rho(s,t) + \rho^2(s,t)\frac{d(s)d(t)}{2} & 1 + \frac{d^2(t)}{2} \end{pmatrix}. \tag{11}$$

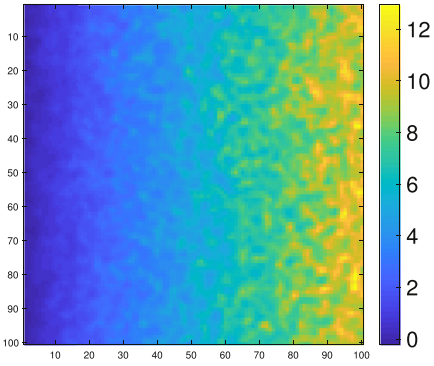
Therefore, the limiting field of the Cohen's d estimator $\hat{d}(s)$ is asymptotically normal with asymptotic variance $1 + \frac{d^2(s)}{2}$. As alluded to in the previous section, it is notable that unlike the sample mean, the asymptotic variance and spatial correlation of the Cohen's d estimator are dependent on the underlying true effect size. In the upcoming section, we will use these properties to motivate a construction for Cohen's d Confidence Sets.



(a) The true underlying signal $\mu(s) = d(s)$, a linear ramp increasing from 0 to 10 in the x-direction.



(b) The sample mean field $\bar{Y}(s)$



(c) The sample Cohen's d field $\hat{d}(s) = \frac{\bar{Y}(s)}{\hat{\sigma}(s)}$

Fig. 2. Visualizing the differences between the sample mean and sample Cohen's d field. For $N = 60$ subjects, we simulated a signal-plus-noise model where the true underlying mean signal $\mu(s)$ was a linear ramp increasing from 0 to 10 across the region (a). To each subject we added Gaussian noise with a homogeneous variance, so that the true Cohen's d effect $d(s)$ was equal to the group mean signal $\mu(s)$. While the sample mean image $\bar{Y}(s)$ is uniformly smooth across the region (b), the sample Cohen's d field $\hat{d}(s)$ becomes rougher from left to right (c).

2.3. *Confidence Sets for Cohen's* d *effect size images*

Once again, consider the model outlined at the start of Section 2.1. For clarity, we reiterate that the spatial CSs for the raw %BOLD change field $\mu(s)$ of focus in our previous work took the form of (5), where k was determined via a Wild t-Bootstrap procedure. This construction of the CSs was motivated by the limiting properties of the field

$$M(s) = \sqrt{N} \cdot \frac{\bar{Y}(s) - \mu(s)}{\hat{\sigma}(s)}. \tag{12}$$

In particular, letting $\partial\mathcal{A}_{c,\mu}$ denote the boundary of $\mathcal{A}_{c,\mu}$ defined in (4), then for a neighbourhood U of $\partial\mathcal{A}_{c,\mu}$, it is assumed in *SSS* that $M(s)$ converges weakly to a smooth Gaussian field $G(s)$ on U with mean zero, unit variance, and with the same (unknown) spatial correlation as each of the ϵ_i.

In the previous section, for the Gaussian one-sample model we derived the convergence in distribution of the function

$$N(s) = \sqrt{N} \cdot \frac{\hat{d}(s) - d(s)}{\sqrt{1 + \frac{d^2(s)}{2}}} \tag{13}$$

to a Gaussian field $\mathcal{G}(s)$ with mean zero, unit variance, and covariance structure

$$\mathrm{Cov}[\mathcal{G}(s), \mathcal{G}(t)] = \frac{\rho(s,t) + \rho(s,t)^2 \frac{d(s)d(t)}{2}}{\sqrt{\left(1 + \frac{d(s)^2}{2}\right)\left(1 + \frac{d(t)^2}{2}\right)}}. \tag{14}$$

This suggests a natural analog to the construction of CSs in (5) for the Cohen's d effect size given by

$$\hat{\mathcal{A}}^+_{c,d} := \left\{ \mathbf{s} : \hat{d}(\mathbf{s}) \geq c + \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(\mathbf{s})}{2}} \right\}, \hat{\mathcal{A}}^-_{c,d} := \left\{ \mathbf{s} : \hat{d}(\mathbf{s}) \geq c - \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(\mathbf{s})}{2}} \right\}. \tag{15}$$

Ideally, we wish to apply the same Wild t-Bootstrap procedure described in Section 2.2 of *BTSN* to approximate the limiting field $\mathcal{G}$ in order to determine k. However, we will now show that such an approach is not viable for Cohen's d, before proposing a modified procedure to solve the problem. Going forward our focus will primarily be on the Cohen's d effect size, and thus for brevity, we will drop the subscript from our notation and refer to the Cohen's d CSs above as $\hat{\mathcal{A}}^+_c$ and $\hat{\mathcal{A}}^-_c$ respectively.

2.4. *Modified residuals for the Cohen's* d *wild* t*-bootstrap*

In *SSS*, it was shown that the limiting coverage of the CSs for the %BOLD effect size $\mu(s)$ is governed by the maximum distribution of the limiting Gaussian field $G(s)$ on the boundary $\partial\mathcal{A}_{c,\mu}$, such that

$$\lim_{n\to\infty} P\left[\hat{\mathcal{A}}^+_{c,\mu} \subset \mathcal{A}_{c,\mu} \subset \hat{\mathcal{A}}^-_{c,\mu}\right] = P\left[\sup_{\mathbf{s}\in\partial\mathcal{A}_{c,\mu}} |G(\mathbf{s})| \leq k\right]. \tag{16}$$

Since the limiting Gaussian field $G(s)$ is unknown, in *BTSN* we implemented a Wild t-Bootstrap procedure to approximate $G(s)$ on the boundary $\partial\mathcal{A}_{c,\mu}$. Defining the standardized residuals,

$$\tilde{\epsilon}_i(s) = \frac{Y_i(s) - \bar{Y}(s)}{\hat{\sigma}(s)}, \tag{17}$$

the Wild t-Bootstrap approximating field is given by

$$\tilde{G}^*(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{\epsilon}_i(s)}{\hat{\sigma}^*(s)}, \tag{18}$$

where the r_i^* are i.i.d. Rademacher variables (i.e. each r_i^* takes the value of -1 or 1 with probability 1/2), and $\hat{\sigma}^*(s)$ is the standard deviation of the current realization of bootstrapped residuals $r_i^* \tilde{\epsilon}_i(s)$. The asterisk $(*)$ indicates that $\tilde{G}^*(s)$ is one of many bootstrap samples; in practice, we would obtain a large number B of bootstrap samples $\tilde{G}^*(s)$, and approximate k as the $(1-\alpha)100$ percentile of the B suprema $\sup_{s \in \partial \hat{A}_c} |\tilde{G}^*(s)|$.

While this method is valid in regards to %BOLD change, for Cohen's d we demonstrate that asymptotically the correlation structure of the approximating field $\tilde{G}^*(s)$ is incorrect. Consider again the Gaussian model in Section 2.2. In this instance, the covariance of the approximating field is:

$$\begin{aligned} \mathrm{Cov}[\tilde{G}^*(s), \tilde{G}^*(t)] &= \frac{1}{N} \sum_{i,j=1}^{N} \frac{\tilde{\epsilon}_i(s)\tilde{\epsilon}_j(t)}{\hat{\sigma}^*(s)\hat{\sigma}^*(t)} \mathrm{Cov}[r_i, r_j] \\ &= \frac{1}{N} \cdot \frac{1}{\hat{\sigma}^*(s)\hat{\sigma}^*(t)} \cdot \frac{1}{\hat{\sigma}(s)\hat{\sigma}(t)} \sum_{i=1}^{N} (Y_i(s) - \bar{Y}(s))(Y_i(t) - \bar{Y}(t)) \\ &\xrightarrow{a.s.} \rho(s,t), \end{aligned} \tag{19}$$

where we note that since the standardized residuals $\tilde{\epsilon}_i(s)$ are asymptotically Gaussian with unit variance, the bootstrap estimate of the standard deviation of the standardized residuals $\hat{\sigma}^*(s)$ converges almost surely to 1. Since (19) and (14) do not agree except for the complete null case of $d(s) = 0$ everywhere, we conclude that the covariance of the approximating field does not converge to the true covariance of the limiting field $\mathcal{G}(s)$.

To solve this problem, we implement a Taylor expansion transformation recently proposed in Telschow et al. (2020) to construct modified residuals with the desired limiting properties. Motivated by the delta method procedures used in Section 2.2, an estimation of the residual field for a single subject i is given by:

$$\mathcal{E}_i(\mathbf{s}) = \frac{Y_i(\mathbf{s})}{Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} = f\left(Y_i(\mathbf{s}), \left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2\right) - f\left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right), \tag{20}$$

where $f(x,y) = \frac{x}{\sqrt{y}}$. A first-order Taylor expansion of $f(x,y)$ about the point $\left(\hat{\mu}(s), \hat{\sigma}^2(s)\right)$ yields the approximating Cohen's d residuals:

$$\begin{aligned} R_i(\mathbf{s}) &= \nabla f\left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right) \left(\left(Y_i(\mathbf{s}), \left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2\right) - \left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right) \right)^{\intercal} \\ &= \frac{Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{2\hat{\sigma}(\mathbf{s})} \left(\frac{\left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2}{\hat{\sigma}^2(\mathbf{s})} - 1 \right). \end{aligned} \tag{21}$$

Normalizing by the estimated standard deviation of the limiting field $\mathcal{G}(s)$, we obtain the modified standardized residuals:

$$\tilde{R}_i(s) = \frac{R_i(s)}{\sqrt{1 + \frac{\hat{d}^2(s)}{2}}}. \tag{22}$$

In Telschow et al. (2020), it is shown that the limiting covariance of $\tilde{R}_i(s)$ is equal to the covariance function of $\mathcal{G}(s)$. Therefore, a modification of (18) leads us to the Cohen's d version of the Wild t-Bootstrap approximating field,

$$\tilde{G}^*(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(s)}{\hat{\sigma}^*(s)}, \tag{23}$$

where now $\hat{\sigma}^*(s)$ is the standard deviation of the bootstrapped Cohen's d residuals $r_i^* \tilde{R}_i(s)$.

While normalization of the $R_i(s)$ by an estimator of the standard deviation of $\mathcal{G}(s)$ provides us with residuals that have the correct limiting properties, for application to fMRI data we wish to optimize the bootstrap in smaller sample sizes. In this regard, it may be preferable to standardize the $R_i(s)$ using an estimator tailored to the sample. Noting that the sample mean of the approximating residuals, $\bar{Y}_R(s) = \frac{1}{N}\sum_{i=1}^{N} R_i(s)$, is equal to zero for all N, letting

$$\hat{\sigma}_R^2(s) = \frac{1}{N} \sum_{i=1}^{N} R_i^2(s), \tag{24}$$

then an alternative to (22) is to normalize the Cohen's d residuals by their sample standard deviation so that

$$\tilde{R}_i(s) = \frac{R_i(s)}{\hat{\sigma}_R(s)}. \tag{25}$$

These standardized residuals can then be used for the Wild t-Bootstrap approximating field given in (23). In this case, the sample standard deviation should also be accounted for in the formation of the CSs, suggesting an alternate construction to (15) given by

$$\hat{A}_c^+ := \left\{ \mathbf{s} : \hat{d}(\mathbf{s}) \geq c + \frac{k}{\sqrt{N}} \hat{\sigma}_R(\mathbf{s}) \right\}, \quad \hat{A}_c^- := \left\{ \mathbf{s} : \hat{d}(\mathbf{s}) \geq c - \frac{k}{\sqrt{N}} \hat{\sigma}_R(\mathbf{s}) \right\}. \tag{26}$$

In Section 3.1, we assess the performance of the CSs on synthetic data when the residuals are standardized using either the estimated limiting variance $\sqrt{1 + \frac{\hat{d}^2(s)}{2}}$ or the sample standard deviation $\hat{\sigma}_R(s)$.

2.5. *Finite properties of the Cohen's* d *estimator and a variance-stabilizing transformation*

Up to now, we have motivated two possible constructions for Cohen's d CSs ((15) and (26)) using the limiting properties of the Cohen's d estimator. Here, we will draw our attention to the distributional properties of $\hat{d}(s)$ for finite samples to make further improvements on these methods, and introduce another novel procedure for obtaining CSs based on Gaussianizing the distribution of $\hat{d}(s)$.

Again, assuming the Gaussian model described in Section 2.2, observing that the Cohen's d estimator can be expressed in the form,

$$\hat{d} = \frac{\bar{Y}}{\hat{\sigma}} = \frac{1}{\sqrt{N}} \cdot \frac{\bar{Y}}{\hat{\sigma}/\sqrt{N}} = \frac{1}{\sqrt{N}} \cdot \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{N}} + \frac{\mu}{\sigma/\sqrt{N}}}{\sqrt{\left(\frac{\hat{\sigma}^2}{\sigma^2/(N-1)}\right)/(N-1)}}, \tag{27}$$

from the RHS of the equality we deduce that $\sqrt{N}\hat{d}$ is characterized by a noncentral t-distribution with noncentrality parameter $\sqrt{N}d$ and $N-1$ degrees of freedom.

Letting

$$C_N = \sqrt{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}, \tag{28}$$

where Γ denotes the gamma function, then the expectation of $\sqrt{N}d$ is given by

$$\mathbb{E}\left[\sqrt{N}\hat{d}\right] = \sqrt{N} C_N d. \tag{29}$$

Therefore, unlike the sample mean, the Cohen's d estimator is biased. To improve the performance of the CSs for small sample sizes, we will account for this bias in the formulation of the CSs. A well-known approximation of C_N is

$$C_N \approx \left(1 - \frac{3}{4N-5}\right)^{-1}. \tag{30}$$

Therefore, a bias-corrected version of the CS construction in (15) given by

$$\hat{A}_c^{\pm} := \left\{ s : \hat{d}(s) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(s)}{2}} \right\}, \tag{31}$$

and similarly, a bias-corrected version of the alternate construction in (26) given by

$$\hat{\mathcal{A}}_c^{\pm} := \left\{ s : \hat{d}(s) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}}\hat{\sigma}_R(s) \right\}. \tag{32}$$

In addition to the formation of the CSs, for any application to real data, the Wild t-Bootstrap described in Section 2.4 must be applied over an approximation of the boundary $\partial\mathcal{A}_c = \{s \in S : d(s) = c\}$. Taking into consideration the bias of the Cohen's d estimator, we will use the plug-in boundary:

$$\partial\hat{\mathcal{A}}_c = \left\{ s \in S : \hat{d}(s) = c\left(1 - \frac{3}{4N-5}\right)^{-1} \right\}. \tag{33}$$

The noncentral t-distribution is asymmetric unless $\mu = 0$; in general, the size of the asymmetry scales with the magnitude of the noncentrality parameter and is inversely proportional to the degrees of freedom. Therefore, we expect the distribution of the Cohen's d estimator to be highly skewed when the true effect size is large and the sample size is small. This conflicts with the symmetric construction of the upper and lower CSs $\hat{\mathcal{A}}_c^+$ and $\hat{\mathcal{A}}_c^-$ given in (31) and (32), suggesting that the coverage performance of these two methods may decline in such situations.

To account for skewness, we adapt a method originally proposed in Laubscher (1960) to stabilize the variance of the noncentral t, transforming to a distribution which is approximately Gaussian, and hence, symmetric. Letting

$$\begin{aligned}
\alpha^* &= \left(\frac{N(8N^2 - 17N + 11)}{(N-3)(4N-5)^2}\right)^{-\frac{1}{2}}, \\
\beta^* &= \left(\frac{N(N-1)}{N-3}\right)^{\frac{1}{2}} \left(\frac{8N^2 - 17N + 11}{(N-3)(4N-5)^2}\right)^{-\frac{1}{2}}, \\
b^* &= \left(\frac{N(8N^2 - 17N + 11)}{(N-3)(4N-5)^2}\right)^{\frac{1}{2}},
\end{aligned} \tag{34}$$

in Appendix A we show that the variance-stabilizing transformation of $\hat{d}$ is given by:

$$\begin{aligned}
\zeta\left(\hat{d}\right) = \sqrt{N}\Bigg[&\alpha^* arcsinh\left(\beta^* \hat{d}\right) - \alpha^* arcsinh\left(\beta^* d\left(1 - \frac{3}{4N-5}\right)^{-1}\right) \\
&+ \frac{1}{2N} b^{*2} \left(d\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2 - 17N + 11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}\Bigg].
\end{aligned} \tag{35}$$

Numerical work presented in Laubscher (1960) shows that the 90th percentile value of $\zeta\left(\hat{d}\right)$ in (35) closely estimates $\phi^{-1}(0.9)$ for a range of true effect sizes d when the sample size is larger than 40, suggesting that – for moderate sample sizes – the distribution of $\zeta\left(\hat{d}\right)$ is approximately Gaussian.

By the monotonicity of the mapping $x \mapsto \alpha^* arcsinh\left(\beta^* x\right)$, the variance-stabilizing transformation provides a further possibility for constructing the CSs in the transformed space $\zeta\left(\hat{d}(S)\right)$. Reconstructing (35), the transformed CSs are given by:

$$\begin{aligned}
\hat{\mathcal{A}}_c^{\pm} = \Bigg\{ s : \alpha^* arcsinh\left(\beta^* \hat{d}(s)\right) \geq \alpha^* arcsinh\left(\beta^* c\left(1 - \frac{3}{4N-5}\right)^{-1}\right) \\
- \frac{1}{2N} b^{*2}\left(c\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nc^2\left(\frac{8N^2 - 17N + 11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}} \pm \frac{k}{\sqrt{N}} \Bigg\}.
\end{aligned} \tag{36}$$

In this case, the Cohen's d residuals given in (21) for the Wild t-Bootstrap must also be modified. An estimation of the transformed residual field for a single subject i is given by

$$\begin{aligned}
\mathcal{E}_i(\mathbf{s}) &= \alpha^* arcsinh\left(\beta^* \frac{Y_i(\mathbf{s})}{Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})}\right) - \alpha^* arcsinh\left(\beta^* \frac{\overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})}\right) \\
&= g\left(Y_i(\mathbf{s}), \left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2\right) - g\left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right),
\end{aligned} \tag{37}$$

where the function $g(x, y) = \alpha^* arcsinh\left(\beta^* \frac{x}{\sqrt{y}}\right)$. Similarly to the methods applied in Section 2.4, a first-order Taylor expansion of $g(x, y)$ about the point $(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s}))$ obtains the transformed Cohen's d residuals

$$\begin{aligned}
\tilde{R}_i(\mathbf{s}) &= \nabla g\left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right)\left(\left(Y_i(\mathbf{s}), \left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2\right) - \left(\overline{Y}(\mathbf{s}), \hat{\sigma}^2(\mathbf{s})\right)\right)^{\intercal} \\
&= \frac{\alpha^* \beta^*}{\sqrt{1 + \beta^{*2}\frac{\overline{Y}^2(\mathbf{s})}{\hat{\sigma}^2(\mathbf{s})}}} \left(\frac{Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{2\hat{\sigma}(\mathbf{s})}\left(\frac{\left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2}{\hat{\sigma}^2(\mathbf{s})} - 1\right)\right).
\end{aligned} \tag{38}$$

In practice, the critical value k in (36) is computed by applying the Wild t-Bootstrap over $\partial\hat{\mathcal{A}}_c$ using the transformed Cohen's d residuals given above for the bootstrap approximating field in (23).

For coherence, we will now formalize the complete procedures to obtain CSs for each of our three proposed CS constructions in (31), (32), and (36).

2.6. Three algorithms for computing Cohen's d CSs

Based on our derivations up to this point, we give three algorithms to compute Cohen's d CSs for data modelled within the one-sample model in Section 2.1. While the first two algorithms are similar, the key difference separating these methods is the estimator of the variance used in the formation of the CSs and for standardizing the Cohen's d residuals. We first describe Algorithm 1, where we use $1 + \frac{\hat{d}^2(s)}{2}$ as the estimator of the variance, motivated by the variance of the limiting field $\mathcal{G}(s)$ derived in Sections 2.2 and 2.3.

Algorithm 1. For observations $Y_1(s), \ldots, Y_N(s)$ modelled by the one-sample linear model in (1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\mathbf{s}) = \frac{\overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.

1. For each observation, $Y_i(s)$, let $\hat{\epsilon}_i(s)$ denote the residual field, $\hat{\epsilon}_i(\mathbf{s}) = Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})$. Then compute the Cohen's d residuals as
$$R_i(\mathbf{s}) = \frac{\hat{\epsilon}_i(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{2\hat{\sigma}(\mathbf{s})}\left(\frac{\hat{\epsilon}_i^2(\mathbf{s})}{\hat{\sigma}^2(\mathbf{s})} - 1\right).$$
2. Normalize the Cohen's d residuals by the estimated limiting standard deviation of the Cohen's d image to obtain the standardized residuals,
$$\tilde{R}_i(\mathbf{s}) = \frac{R_i(\mathbf{s})}{\sqrt{1 + \frac{\overline{Y}^2(\mathbf{s})}{2\hat{\sigma}^2(\mathbf{s})}}}.$$
3. Draw N i.i.d. Rademacher variables $r_1^*, \ldots, r_N^*$, and compute the Wild t-Bootstrap approximating field,
$$\tilde{G}^*(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(s)}{\hat{\sigma}^*(s)},$$
where $\hat{\sigma}^*(s)$ is the bootstrap standard deviation of the bootstrapped residuals $r_i^* \tilde{R}_i(s)$.
4. Obtain the value $k^* = \sup_{s \in \partial\hat{\mathcal{A}}_c} |G^*(s)|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{ s \in S : \hat{d}(s) = c\left(1 - \frac{3}{4N-5}\right)^{-1} \right\}$.
5. For a large number of bootstrap replications B repeat steps 3. and 4., obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, \ldots, k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.
6. Obtain the Cohen's d CSs,
$$\hat{\mathcal{A}}_c^{\pm} := \left\{ s : \hat{d}(s) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}}\sqrt{1 + \frac{\hat{d}^2(s)}{2}} \right\}.$$

For Algorithm 2, we use the sample variance of the Cohen's d residuals $\hat{\sigma}_R^2(s)$ as the variance estimator, motivated by our workings in Section 2.4.

Algorithm 2. For observations $Y_1(s), \dots, Y_N(s)$ modelled by the one-sample linear model in (1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\mathbf{s}) = \frac{\overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.

1. For each observation, $Y_i(s)$, let $\hat{\epsilon}_i(s)$ denote the residual field, $\hat{\epsilon}_i(\mathbf{s}) = Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})$. Then compute the Cohen's d residuals as
$$R_i(\mathbf{s}) = \frac{\hat{\epsilon}_i(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{2\hat{\sigma}(\mathbf{s})}\left(\frac{\hat{\epsilon}_i^2(\mathbf{s})}{\hat{\sigma}^2(\mathbf{s})} - 1\right).$$
2. Normalize the Cohen's d residuals by their sample standard deviation to obtain the standardized residuals,
$$\tilde{R}_i(s) = \frac{R_i(s)}{\hat{\sigma}_R(s)}.$$
3. Draw N i.i.d. Rademacher variables $r_1^*, \dots, r_N^*$, and compute the Wild t-Bootstrap approximating field,
$$\tilde{G}^*(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(s)}{\hat{\sigma}^*(s)},$$
where $\hat{\sigma}^*(s)$ is the bootstrap standard deviation of the bootstrapped residuals $r_i^* \tilde{R}_i(s)$.
4. Obtain the value $k^* = \sup_{s \in \partial\hat{\mathcal{A}}_c} |G^*(s)|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{ s \in S : \hat{d}(s) = c\left(1 - \frac{3}{4N-5}\right)^{-1} \right\}$.
5. For a large number of bootstrap replications B repeat steps 3. and 4., obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, \dots, k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.
6. Obtain the Cohen's d CSs,
$$\hat{\mathcal{A}}_c^{\pm} := \left\{ s : \hat{d}(s) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}} \hat{\sigma}_R(s) \right\}.$$

Finally, Algorithm 3 is based on the derivations in Section 2.5, transforming the estimated Cohen's d image to a field which is approximately Gaussian. This is done to stabilize the variance and remove the skew of the Cohen's d estimator, which may adversely effect the performance of the CSs. By the monotonicity of the transformation, the CSs obtained using this method are valid for inference on the true (un-transformed) Cohen's d effect size.

Algorithm 3. For observations $Y_1(s), \dots, Y_N(s)$ modelled by the one-sample linear model in (1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\mathbf{s}) = \frac{\overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.

1. For each observation, $Y_i(s)$, let $\hat{\epsilon}_i(s)$ denote the residual field, $\hat{\epsilon}_i(\mathbf{s}) = Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})$. Then compute the transformed, variance-stabilized Cohen's d residuals as
$$\tilde{R}_i(\mathbf{s}) = \frac{\alpha^*\beta^*}{\sqrt{1 + \beta^{*2}\frac{\overline{Y}^2(\mathbf{s})}{\hat{\sigma}^2(\mathbf{s})}}}\left(\frac{Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})}{\hat{\sigma}(\mathbf{s})} - \frac{\overline{Y}(\mathbf{s})}{2\hat{\sigma}(\mathbf{s})}\left(\frac{\left(Y_i(\mathbf{s}) - \overline{Y}(\mathbf{s})\right)^2}{\hat{\sigma}^2(\mathbf{s})} - 1\right)\right).$$
2. Draw N i.i.d. Rademacher variables $r_1^*, \dots, r_N^*$, and compute the Wild t-Bootstrap approximating field,
$$\tilde{G}^*(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(s)}{\hat{\sigma}^*(s)},$$
where $\hat{\sigma}^*(s)$ is the bootstrap standard deviation of the bootstrapped residuals $r_i^* \tilde{R}_i(s)$.
3. Obtain the value $k^* = \sup_{s \in \partial\hat{\mathcal{A}}_c} |G^*(s)|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{ s \in S : \hat{d}(s) = c\left(1 - \frac{3}{4N-5}\right)^{-1} \right\}$.
4. For a large number of bootstrap replications B repeat steps 3 and 4, obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, \dots, k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.
5. Obtain the Cohen's d CSs,
$$\hat{\mathcal{A}}_c^{\pm} = \left\{ s : \alpha^* arcsinh\left(\beta^* \hat{d}(s)\right) \geq \alpha^* arcsinh\left(\beta^* c\left(1 - \frac{3}{4N-5}\right)^{-1}\right) - \frac{1}{2N} b^{*2}\left(c\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nc^2\left(\frac{8N^2 - 17N + 11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}} \pm \frac{k}{\sqrt{N}} \right\}. \quad (39)$$

3. Methods

3.1. Simulation setup

In this section we describe the settings used to evaluate the performance of each of the three algorithms for obtaining Cohen's d CSs on synthetic data. For each method, we simulate 3000 independent samples of the Gaussian one-sample model

$$Y_i(s) = \mu(s) + \epsilon_i(s), \qquad i = 1, \dots, N$$

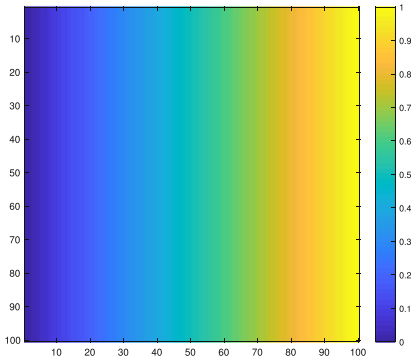
using a range of signals $\mu(s)$, Gaussian noise structures $\epsilon_i(s)$ with stationary and non-stationary variance $\sigma^2(s)$, in two- and three-dimensional regions S. To compute the critical value k, we apply the given method's Wild t-Bootstrap procedure with $B = 5000$ bootstrap samples on the estimated boundary $\partial\hat{\mathcal{A}}_c$ that must be used for application to real data. We obtain the boundary using the linear interpolation method described in Section 2.3 of *BTSN*. We then compute the empirical coverage using the interpolation assessment method described in Section 2.4 of *BTSN*, given as the percentage of trials for which the true thresholded signal $\mathcal{A}_c$ contains the upper CS $\hat{\mathcal{A}}_c^+$ and is contained within the lowers CS $\hat{\mathcal{A}}_c^-$ (i.e. the proportion of trials for which $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$). In each simulation, we apply the given method to sample sizes of $N = 30, 60, 120, 240$ and 480, and for each of the three nominal coverage probability levels $1 - \alpha = 0.80, 0.90$ and 0.95.

3.2. 2D simulations

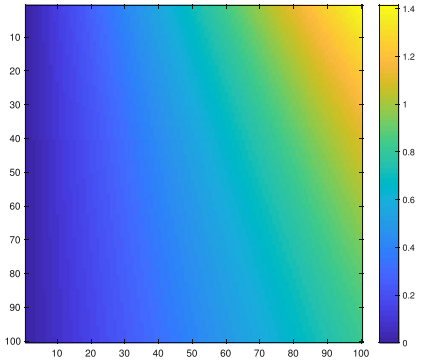
We analyzed the performance of the three algorithms to obtain Cohen's d CSs on a square region of size 100×100. For the true underlying signal $\mu(s)$ we considered two different raw effects: first, a linear ramp that increased from a magnitude of 0 to 1 in the x-direction while remaining constant in the y-direction. Second, a circular effect, created by placing a circular phantom of magnitude 1 and radius 30 in the centre of the search region, which was then smoothed using a 3-voxel FWHM Gaussian kernel. If we were to assume that each voxel had a size of 2 mm^3, we note that this would amount to applying smoothing with a 6 mm FWHM kernel, a fairly typical setting used in fMRI analyses.

To each of these signals we added subject-specific Gaussian noise $\epsilon_i(s)$, obtained from smoothing white noise with a 3 voxel FWHM Gaussian kernel, with homogeneous and non-homogeneous variance structures: the first noise field had a spatially constant standard deviation of 1, and therefore in this case the true Cohen's d effect was identical to the underlying signal $\mu(s)$. The second field had a linearly increasing standard deviation structure in the y-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$ while remaining constant in the x-direction. Thus, the variance of this noise field spatially increased in the y-direction from 0.5 to 1.5 in a non-linear fashion.

The true Cohen's d fields $d(s)$ for the linear ramp signal with homogeneous and heterogeneous noise are shown in Fig. 3. The corresponding Cohen's d fields for the circular signal are shown in Fig. 4. Altogether, for the three algorithms, the two underlying signals and two noise sources gave us 12 different simulation setups; for all of the simulations, we obtained Cohen's d CSs for the noise-free cluster $\mathcal{A}_c$ at a cluster-forming

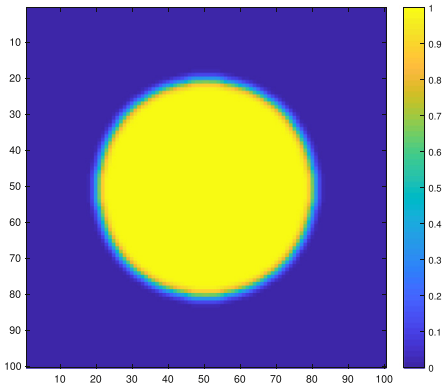


(a) The Cohen's d field $d(\boldsymbol{s})$ for the linear ramp signal $\mu(\boldsymbol{s})$ and homogeneous noise structure.

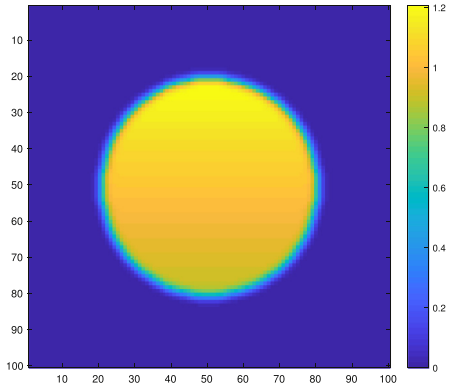


(b) The Cohen's d field $d(\boldsymbol{s})$ for the linear ramp signal $\mu(\boldsymbol{s})$ and heterogeneous noise structure.

Fig. 3. The two Cohen's d effects corresponding to the linear ramp signal $\mu(\boldsymbol{s})$. On the left, the subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ has a spatially constant standard deviation of 1, and therefore $d(\boldsymbol{s}) = \mu(\boldsymbol{s})$. On the right, $\epsilon_i(\boldsymbol{s})$ had a spatially increasing standard deviation structure in the y-direction (from top-to-bottom), while remaining constant in the x-direction.



(a) The Cohen's d field $d(\boldsymbol{s})$ for the circular signal $\mu(\boldsymbol{s})$ and homogeneous noise structure.



(b) The Cohen's d field $d(\boldsymbol{s})$ for the circular signal $\mu(\boldsymbol{s})$ and heterogeneous noise structure.

Fig. 4. The two Cohen's d effects corresponding to the circular signal $\mu(\boldsymbol{s})$. On the left, the subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ has a spatially constant standard deviation of 1, and therefore $d(\boldsymbol{s}) = \mu(\boldsymbol{s})$. On the right, $\epsilon_i(\boldsymbol{s})$ had a spatially increasing standard deviation structure in the y-direction (from top-to-bottom), while remaining constant in the x-direction.

threshold of $c = 0.8$. In Chapter 2.2 of Cohen (2013), $d = 0.8$ was classified as a 'large effect'; for group-level analyses of large-sample fMRI data with ample statistical power (such as the HCP or UK Biobank), effect sizes of this magnitude may be used to assess brain areas where practically significant activation has occurred.

3.3. 3D simulations

Four signal types $\mu(\boldsymbol{s})$ were considered to analyze the performance of the three algorithms in three dimensions. The first three of these signals were generated synthetically on a cubic region of size $100 \times 100 \times 100$: firstly, a small spherical effect, created by placing a spherical phantom of magnitude 1 and radius 5 in the centre of the search region, which was then smoothed using a 3-voxel FWHM Gaussian kernel. Secondly, a larger spherical effect, generated identically to the first effect with the exception that the spherical phantom had a radius of 30. Lastly, we created an effect by placing four spherical phantoms of magnitude 1 in the region of varying radii and then smoothing the entire image using a 3-voxel FWHM Gaussian.

Each of the images were rescaled after smoothing to have a maximum intensity of 1. For the small and large spherical effect an imagewise rescaling was applied, where all locations in the smoothed map were divided through by the maximum intensity across the region. For the final effect, because parts of the four spherical phantoms overlapped after smoothing, the signal intensities in these regions summed to greater than 1. In this case, we reduced the intensities in these areas to have a magnitude of 1 while leaving the rest of the image untouched. This ensured that the signal at the center of each spherical phantom had a magnitude of 1, coinciding with the previous small and large spherical effect signal types (see Fig. 5, the signal at the center of each spherical phantom in plot (c) is 1, in correspondence with the small and large spherical effects in plots (a) and (b)).

Similar to the two-dimensional simulations, for the three signals described above we added white noise smoothed using a 3-voxel FWHM Gaussian kernel with homogeneous and heterogeneous variance structures. The first noise field had a spatially constant standard deviation of 1, while the second field had a linearly increasing standard deviation in the z-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$, while remaining constant in both the x- and y-directions. As demonstrated for the 2D simulations in Figs. 3 and 4, this lead to two different true Cohen's d effect-size images $d(\boldsymbol{s})$ corresponding to the homogeneous and heterogeneous standard deviation fields $\sigma(\boldsymbol{s})$ used for the noise.

For the final signal type, we took advantage of big data that has been made available through the UK Biobank in an attempt to generate an effect that replicated the true %BOLD change induced during an fMRI task. We randomly selected 4000 subject-level contrast of parameter estimate result maps from the Hariri Faces/Shapes task-fMRI data

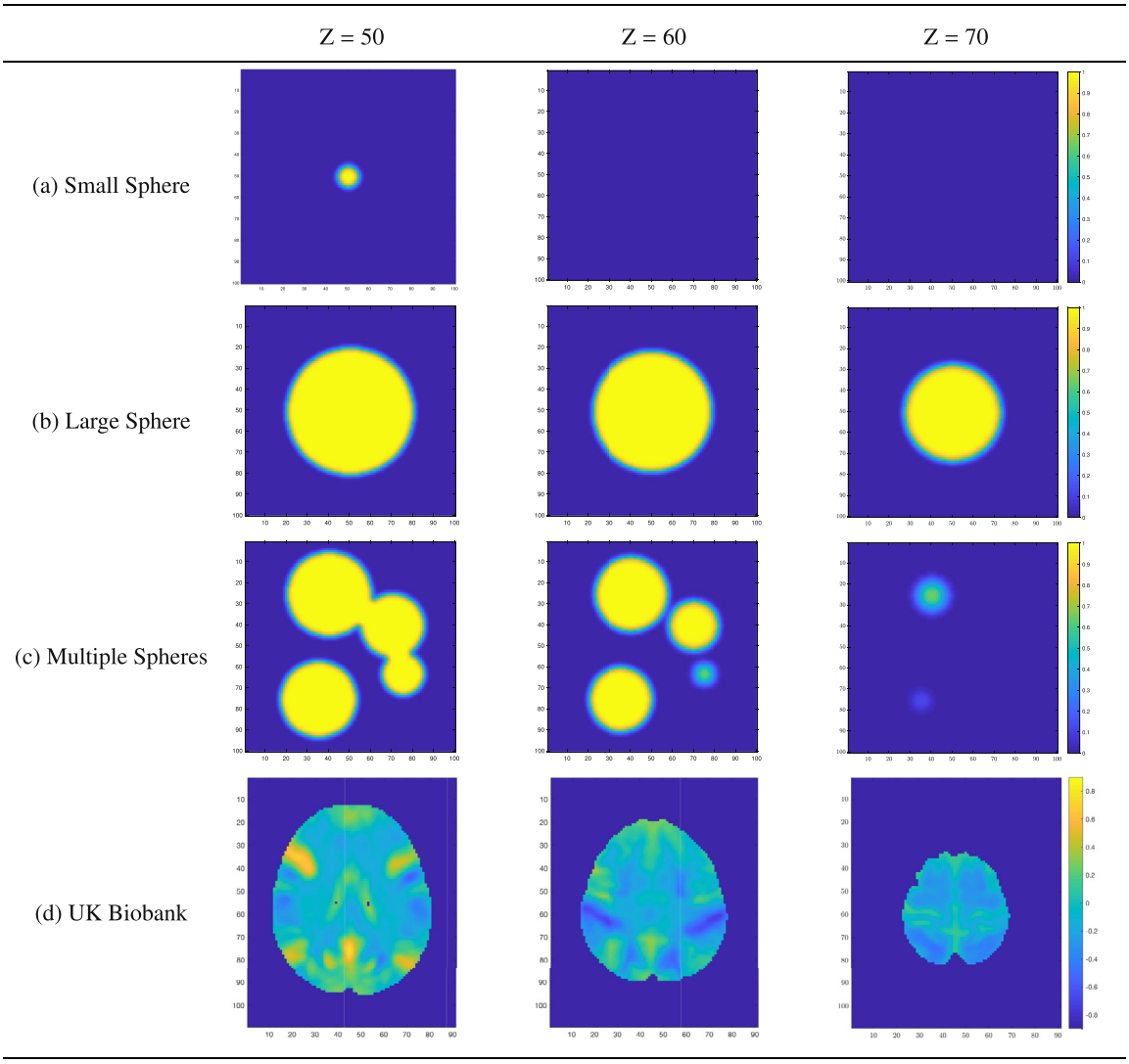


Fig. 5. Four of the Cohen's d fields $d(s)$ used for the 3D simulations. Plots (a)–(c) show the Cohen's d field for the three different spherical effects $\mu(s)$ when Gaussian noise with spatially homogeneous standard deviation was added to the signal. Plot (d) shows the Cohen's d field corresponding to the UK Biobank full mean and standard deviation images. Note that the colormap limits for the first three Cohen's d effect-size images are from 0 to 1, while the colormap limits for the UK Biobank image is from −0.9 to 0.9.

collected as part of the UK Biobank brain imaging study. Full details on how the data were acquired and processed is given in *Miller et al. (2016)*, *Alfaro-Almagro et al. (2018)* and the UK Biobank Showcase; information on the task paradigm is given in *Hariri et al. (2002)*. From these contrast maps, we computed a group-level full mean and full standard deviation image, considering all voxels where at least one hundred subjects had data (instead of discarding voxels with any missing data). In the final simulation, we used the group-level Biobank mean image as the true underlying signal $\mu(s)$ for each subject, and the full standard deviation image was used for the standard deviation of each simulated subject-specific Gaussian noise field $\epsilon_i(s)$ added to the true signal. Because of the considerably large sample size of high-quality data from which these maps have been obtained, we anticipate that both of these images are highly representative of the true underlying fields that they approximate. Both images were masked using an intersection of all 4000 of the subject-level brain masks.

Once again, we smoothed the noise field using a 3-voxel FWHM Gaussian kernel; as the Biobank maps had voxel sizes of 2 mm^3, this equated to applying 6 mm FWHM smoothing to the noise field of the original data.

In Fig. 5, we show the true underlying Cohen's d fields for the three synthetic 3D effects with homogeneous noise structure, and the Cohen's d field corresponding to the UK Biobank full mean and standard deviation (a histogram of the UK Biobank Cohen's d field is provided in Fig. B.14). For all four signal types we obtained Cohen's d Confidence Sets for the threshold $c = 0.8$, and in order to assess if a change of threshold could affect the performance of the CSs, we also obtained Cohen's d CSs using a threshold of $c = 0.5$ for the final UK Biobank signal type.

3.4. Application to Human Connectome Project data

To provide a real-data demonstration of the three methods proposed in this work, we computed Cohen's d CSs on 80 participants data from the Unrelated 80 package released as part of the Human Connectome Project (HCP, S1200 Release) using all three algorithms described in Section 2.6. Cohen's d CSs were obtained for the subject-level 2-back vs 0-back contrast maps from the working memory task results included with the HCP dataset. For a comparison with standard fMRI inference procedures, we also performed a traditional statistical group-level analysis on the data. A one-sample t-test was carried out in SPM, using a

voxelwise FWE-corrected threshold of $p < 0.05$ obtained via permutation test with SPM's SnPM toolbox.

For the working memory task participants were presented with pictures of places, tools, faces and body parts in a block design. The task consisted of two runs, where on each run a separate block was designated for each of the image categories, making four blocks in total. Within each run, for half of the blocks participants undertook a 2-back memory task, while for the other half a 0-back memory task was used. Eight EVs were included in the GLM for each combination of picture category and memory task (e.g. 2-back Place); we compute CSs on the subject-level contrast images for the 2-back vs 0-back contrast results that contrasted the four 2-back related EVs to the four 0-back EVs.

Imaging was conducted on a 3T Siemans Skyra scanner using a gradient-echo EPI sequence; TR = 720 ms, TE = 33.1 ms, 208×180 mm FOV, 2.0 mm slice thickness, 72 slices, 2.0 mm isotropic voxels, and a multi-band acceleration factor of 8. Preprocessing of the subject-level data was carried out using tools from FSL and Freesurfer following the 'fMRIVolume' HCP Pipeline fully described in Glasser et al. (2013). To summarize, the fundamental steps carried out to each individual's functional 4D time-series data were gradient unwarping, motion correction, EPI distortion correction, registration of the functional data to the anatomy, non-linear registration to MNI space (using FSL's Non-linear Image Registration Tool, FNIRT), and global intensity normalization. Spatial smoothing was applied using a Gaussian kernel with a 4 mm FWHM.

Modelling of the subject-level data was conducted with FSL's FMRIB's Improved Linear Model (FILM). The eight working task EVs were included in the GLM, with temporal derivatives terms added as confounds of no interest, and regressors were convolved using FSL's default double-gamma hemodynamic response function. The functional data and GLM were temporally filtered with a high-pass frequency cutoff point of 200 s, and time series were prewhitened to remove autocorrelations from the data.

Mirroring the methods used in *BTSN*, we applied additional smoothing to the final contrast maps to mimic images smoothed using a 6 mm FWHM Gaussian kernel. This is a more typical degree of smoothing applied to functional data than the 4 mm kernel originally used in the HCP analysis pipeline.

4. Results

4.1. 2D simulations

Empirical coverage results for each of the three algorithms are presented for the linear ramp signal in Fig. 6 and for the circular signal in Fig. 7, where in all simulations a Cohen's d threshold of $c = 0.8$ was applied. In both figures, on the top row we display the coverage results obtained when the standard deviation field of the noise was homogeneous across the region (corresponding to Fig. 3(a) for the linear ramp, Fig. 4(a) for the circle), and on the bottom row we display the equivalent

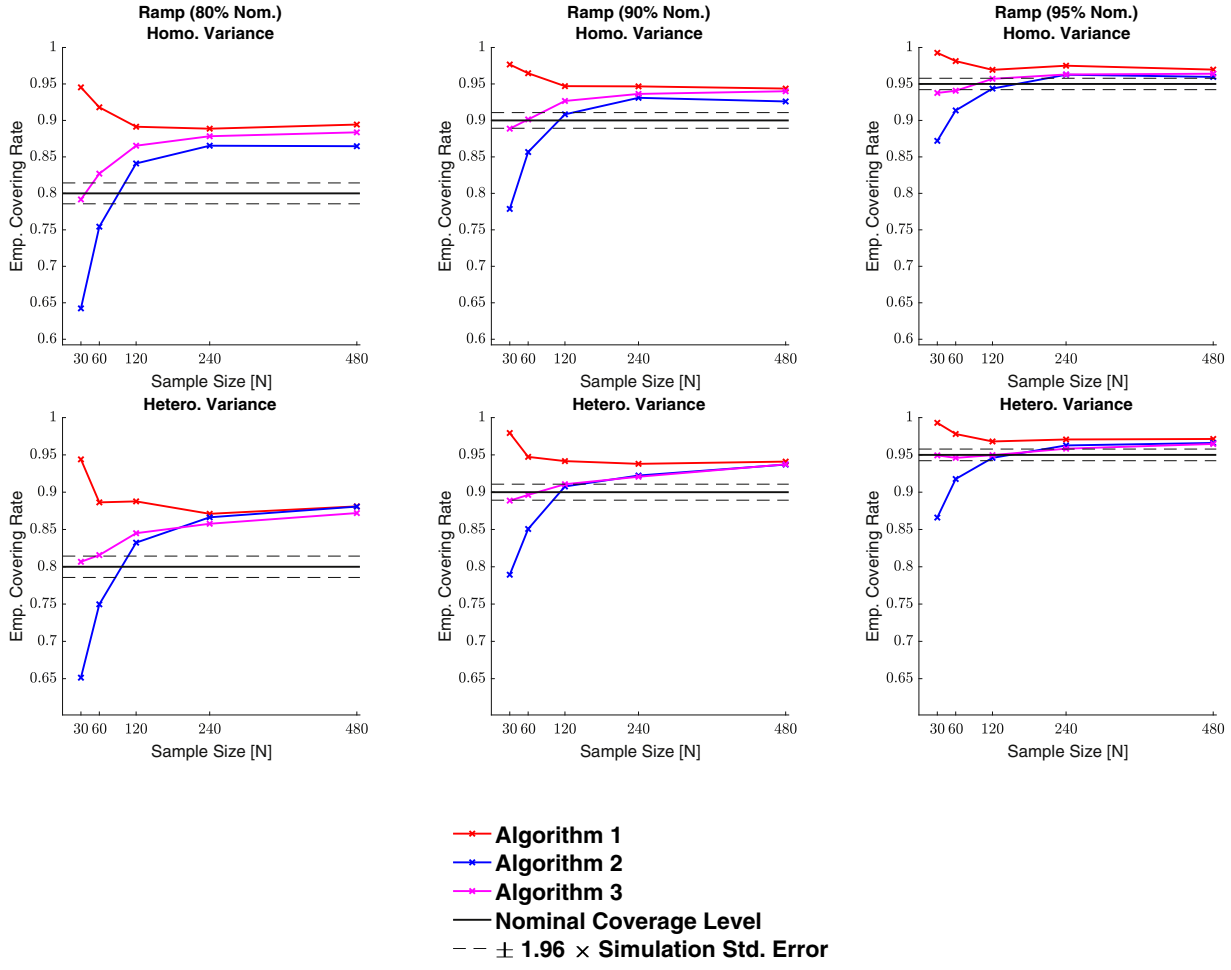


Fig. 6. Coverage results for the linear ramp signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. For large sample sizes the empirical coverage performance of all three algorithms was similar, hovering slightly above the nominal level in all simulations. As the sample size was made smaller the degree of over-coverage became larger for Algorithm 1, while empirical coverage for Algorithm 2 fell below the nominal target. Algorithm 3 performed best, with all results remaining particularly close to the nominal target level for simulations using a 95% confidence level (right plots).

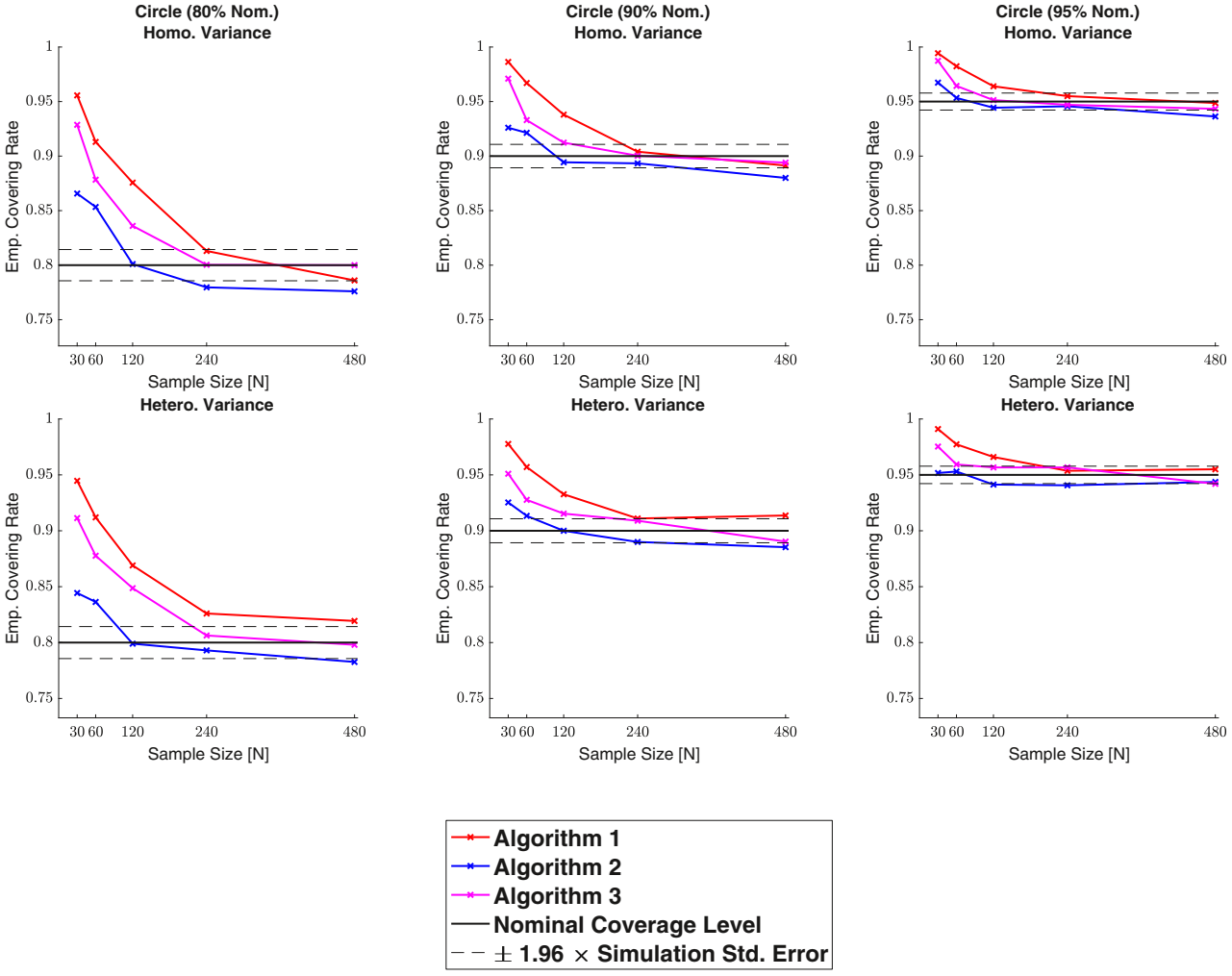


Fig. 7. Coverage results for the circular signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. All algorithms performed well, and unlike the linear ramp, empirical coverage for all three methods converged towards the nominal level. For smaller sample sizes there was a larger degree of over-coverage, most noticeably for simulations using the 80% nominal target. Overall, Algorithm 2 performed marginally better than the other two methods, and Algorithm 1 performed the worst.

results when the standard deviation field was spatially heterogeneous (Fig. 3(a) and Fig. 4(b) for the linear ramp and circle respectively).

For the linear ramp, across all confidence levels $1-\alpha = 0.80, 0.90,$ and 0.95 we observed valid, over-coverage for all three algorithms when sufficiently large sample sizes ($N \geq 60$) were used. In all plots, it appears that the coverage rates for the three algorithms are converging to the same value, slightly above the nominal target. Specifically, for the nominal target level of 80%, in both the homogeneous and heterogeneous cases all empirical results seem to be converging to around 88% (Fig. 6, left-side plots). For the 95% target, the scale of disagreement between the empirical results and the nominal target is smaller; here, all coverage results hover close to 96% for $N = 240$ and 480 (Fig. 6, right-side plots).

While for larger sample sizes the performance of all three algorithms was similar, there was greater disparity between the methods for simulations using the smaller sample sizes of $N = 30$ and 60. Here, the empirical coverage results for Algorithm 1 were consistently higher than the other two methods, with the degree of over-coverage increasing as the sample size was lowered. On the other hand, the coverage results for Algorithms 2 and 3 both decreased as the sample size was made smaller; while there was only a slight dip in coverage for Algorithm 3 here, the drop-off for Algorithm 2 was more considerable, with coverage results falling substantially below the nominal target level when $N = 30$.

For the circular signal, on the whole all three methods performed well. In this instance, almost all empirical results for Algorithms 2 and 3 lay within the 95% confidence interval of the nominal coverage rate (blue and magenta curves sandwiched between black dashed lines for all plots in Fig. 7), with Algorithm 2 performing marginally better. While we observed greater over-coverage for the smaller sample sizes, most substantially in simulations using the 80% nominal target (Fig. 7, left-side plots), empirical coverage converged towards the nominal level for all three algorithms.

Finally, the use of homogeneous or heterogeneous noise in the model had minimal impact on any of the algorithm's empirical coverage performance for either of the signals. This is exemplified in Figs. 6 and 7, where in both cases the homogeneous coverage plots presented in the top row are almost identical to the corresponding heterogeneous plots shown below.

4.2. 3D simulations

Empirical coverage results for each of the three algorithms are presented in Figs. 8, 9, 10 and 11 respectively for each of the four 3D signal types displayed in Fig. 5 (small sphere, large sphere, multiple spheres, UK Biobank), where in all simulations a Cohen's d threshold of $c = 0.8$ was applied. In Fig. C.15, results are presented for the UK Biobank sig-

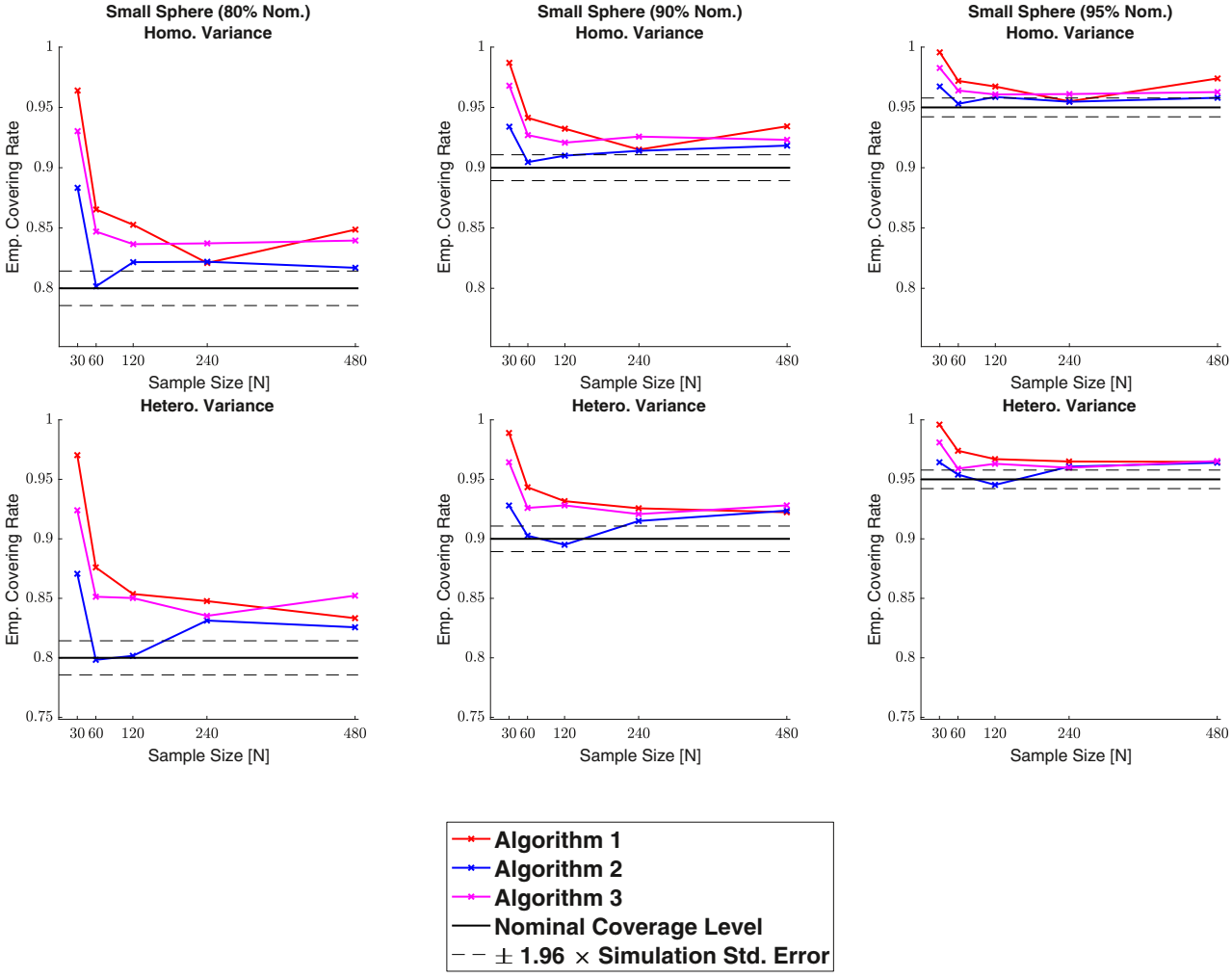


Fig. 8. Coverage results for the small sphere signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. In general, empirical coverage remained above the nominal level across all simulations, and for the 95% confidence level (right plots), the results of all three methods fell close to the nominal target (with some over-coverage for $N = 30$). All methods were robust as to whether the subject-level noise had homogeneous or heterogeneous variance structure. Because of this, there are minimal differences comparing the plots between both rows.

nal type using a smaller Cohen's d threshold of $c = 0.5$. For the spherical effects (Figs. 8–10), on the top row we display the coverage results obtained when the standard deviation field of the noise was homogeneous across the region, and on the bottom row we display the equivalent results when the standard deviation field was spatially heterogeneous. For the UK Biobank signal (Fig. 11), the full standard deviation image computed from the UK Biobank data was used for the standard deviation field of the noise, and hence in this case there is only one row of results.

Across all 3D simulations we observed consistencies between the results obtained with each of the three algorithms: in general, empirical coverage for all methods came above the nominal target, with increasing severity for smaller sample sizes. Similar to the 2D simulations, the extent of over-coverage was smaller when a larger confidence level was used. Comparing the three methods, coverage results for Algorithm 1 were considerably higher than the other two methods, particularly when a small sample size and confidence level were used. For the 'large sphere' and 'multiple spheres' signal types, Algorithm 1 suffered with over-coverage of above 15% in simulations with a sample of size of $N = 30$ and 60 and a nominal target level of 80% (Figs. 9 and 10, left-side plots). For both of these signals, there was still a considerable amount of over-coverage when larger sample sizes of $N = 120, 240$ and (to a lesser degree) 480 were used. On the other hand, Algorithms 2 and 3 performed similarly in large sample sizes across all simulations, with empirical coverage results coming slightly above the nominal target. Notably, both of these algorithms performed very well for simulations with a 95% nominal target level (all figures, right-side plots). Differences between these two methods were more distinguished for smaller sample sizes of $N = 30$ and 60, where we observed a greater degree of over-coverage for Algorithm 3. Consequentially, Algorithm 2's results came closer to the nominal target here, although for the 'multiple sphere' and 'UK Biobank' signal types, in some cases Algorithm 2's results fell slightly *below* the nominal level (Figs. 10 and 11). Overall, empirical coverage for Algorithm 3 was the most uniform of the three methods across moderate and large sample sizes.

Comparing Figs. 8 and 9, we observed a slight deterioration in the performance of all three algorithms when moving from the small sphere signal type to the large sphere. In particular, results obtained from applying the three methods to the large sphere fell further above the nominal target relative to the small sphere. This was most severe for Algorithm 1, where differences between the two sets of results were larger than 10% for the 80% confidence level (Figs. 8 and 9, left-side plots). These dif-

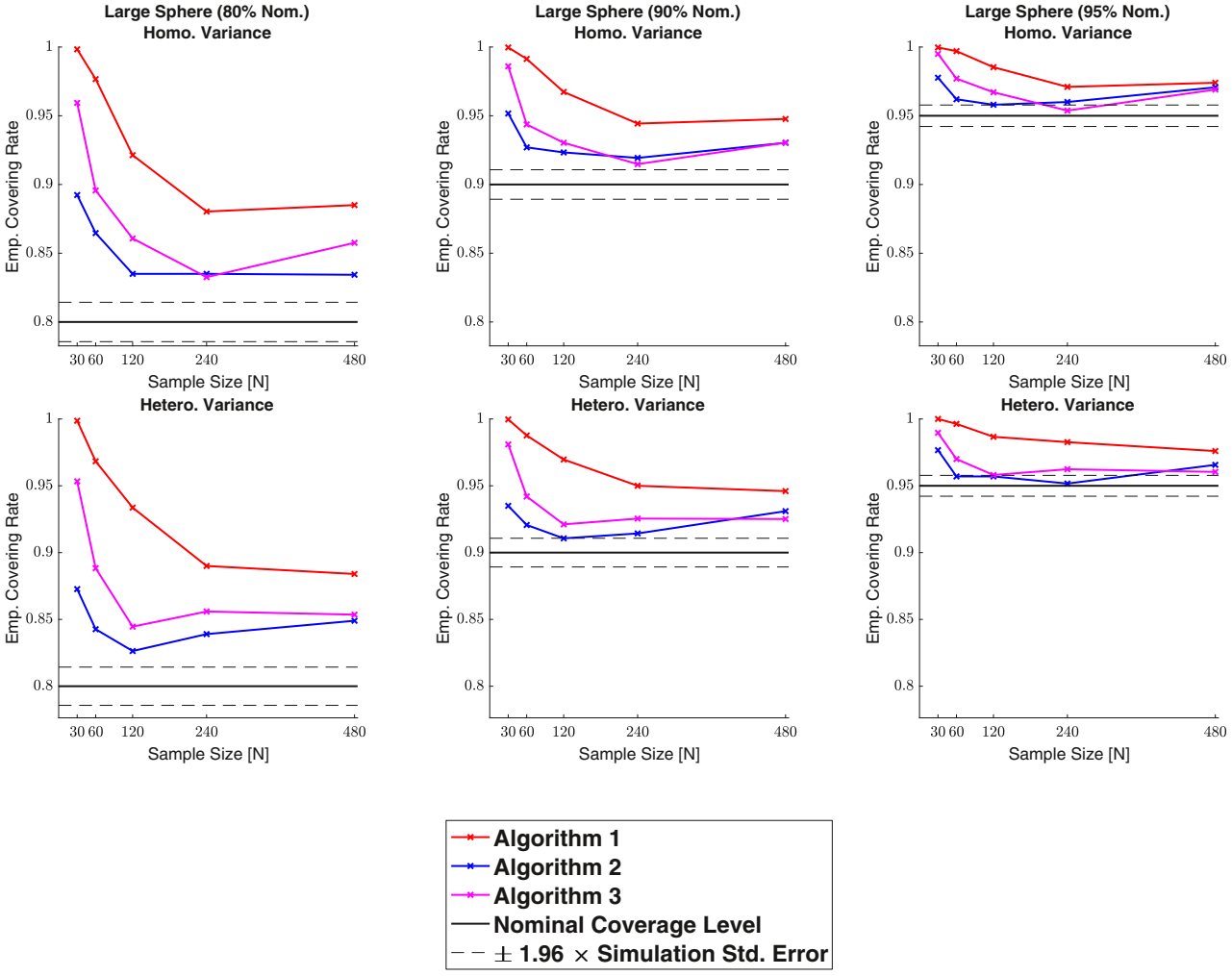


Fig. 9. Coverage results for the large sphere signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. Compared with the small sphere results displayed in Fig. 9, empirical coverage results were higher for all three methods here. Algorithm 1 suffered from a particularly large degree of over-coverage for simulations with a small sample size. Coverage performance for Algorithms 2 and 3 was closer in resemblance to the corresponding small sphere results, with Algorithm 2 performing slightly better. This suggests that both of these methods are fairly robust to changes in the boundary length.

ferences were comparatively smaller for Algorithms 2 and 3, where we observed only a slight increase in empirical coverage, particularly for simulations using larger sample sizes. This would suggest that both of these methods are fairly robust to changes in the boundary length. We observed a similar sort of degradation in performance for the two simulation results obtained using the UK Biobank signal type, where simulations using the smaller threshold of $c = 0.5$ (Fig. C.15) fell further above the nominal target relative to the simulations using a threshold of $c = 0.8$. Once again, this may be attributable to changes in the boundary length between the two simulations (the boundary length was longer for the smaller threshold of $c = 0.5$). Rather than a degradation in the three algorithms' performances per se, we suspect that inaccuracies in the methods used to assess the simulations may have induced a positive bias into the coverage results for signals with a longer boundary (see the end of Section 5.2 for more on this).

Finally, the use of homogeneous or heterogeneous noise in the model once again had very little impact on the performance of all three algorithms. Nevertheless, for simulations with small sample sizes, a heterogeneous noise structure led to a slight decrease in the empirical coverage results for Algorithms 2 and 3 (Figs. 8–10, left-side plots).

4.3. Human Connectome Project

Cohen's d Confidence Sets obtained by applying Algorithm 3 to 80 subjects' contrast data from the Human Connectome Project are shown in Fig. 12. CSs computed on the same data using Algorithm 1 and Algorithm 2 are displayed in Figs. D.1 and D.2 respectively. For each figure, we display the CSs obtained from applying the specified algorithm with three separate thresholds, $c = 0.5, 0.8$, and 1.2. These three Cohen's d effect sizes were classified as medium, large, and very large in Cohen (2013).

In the top plot of Fig. 12, the red upper CSs localized brain regions within the frontal cortex that are commonly associated with working memory. This included areas of the superior frontal gyrus (left and right, all slices), middle frontal gyrus (left, coronal slice), paracingulate gyrus (left and right, axial slice) and insular cortex. Other brain areas encapsulated inside the upper CS were the angular gyrus (left and right, axial slice), cerebellum (left and right, sagittal slice) and precuneus (left and right, axial slice). For all these regions, the method identified clusters of voxels where we can assert with 95% confidence there was a Cohen's d effect size greater than 0.5.

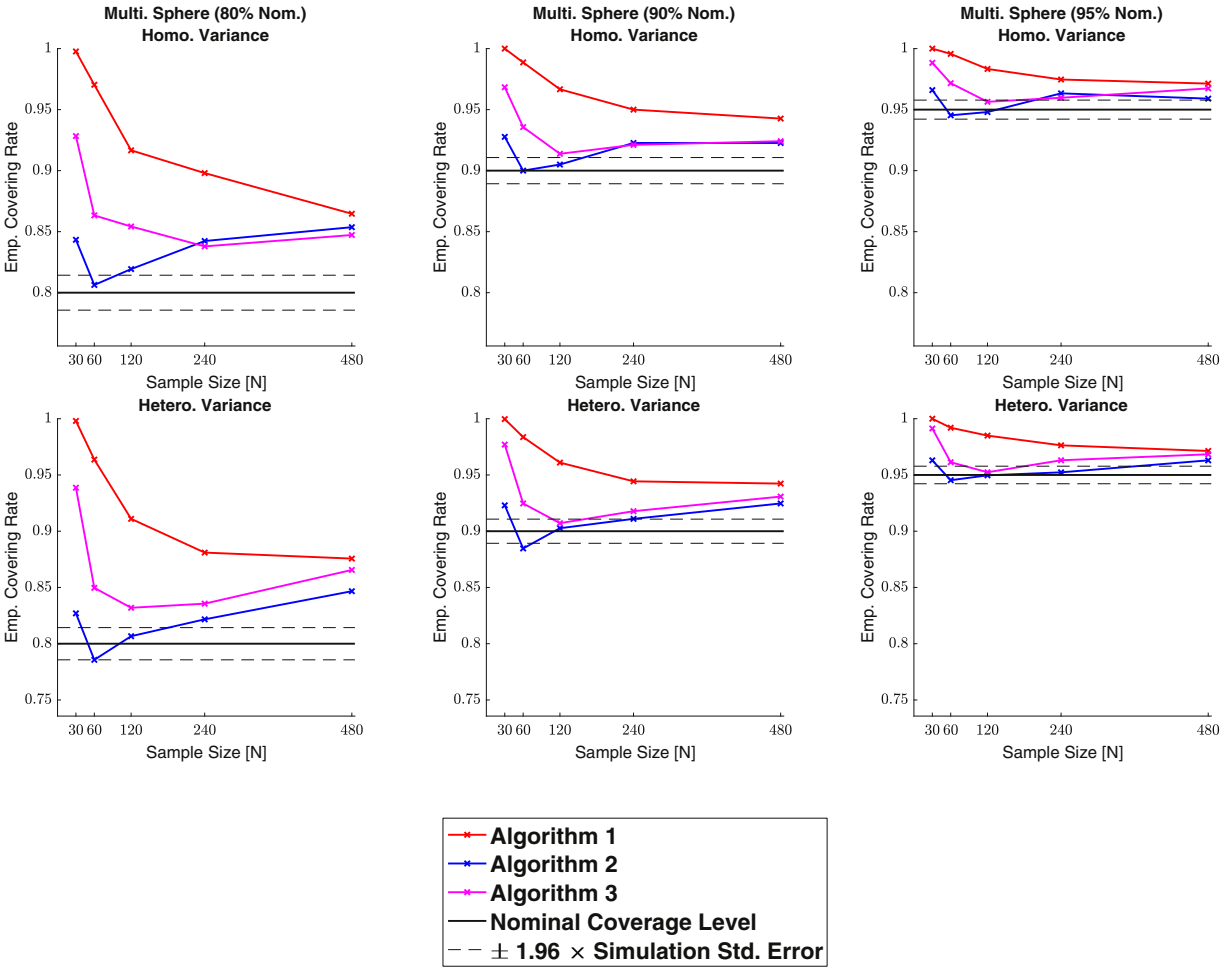


Fig. 10. Coverage results for the multiple spheres signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. Algorithms 2 and 3 both performed well, particularly for the 95% confidence level, where for moderate-to-large sample sizes coverage remained in the vicinity of the 95% confidence interval of the nominal target. Once again, the degree of over-coverage increased as the sample size was made smaller, most severely for Algorithm 1, while Algorithm 2 remained relatively close to the nominal level.

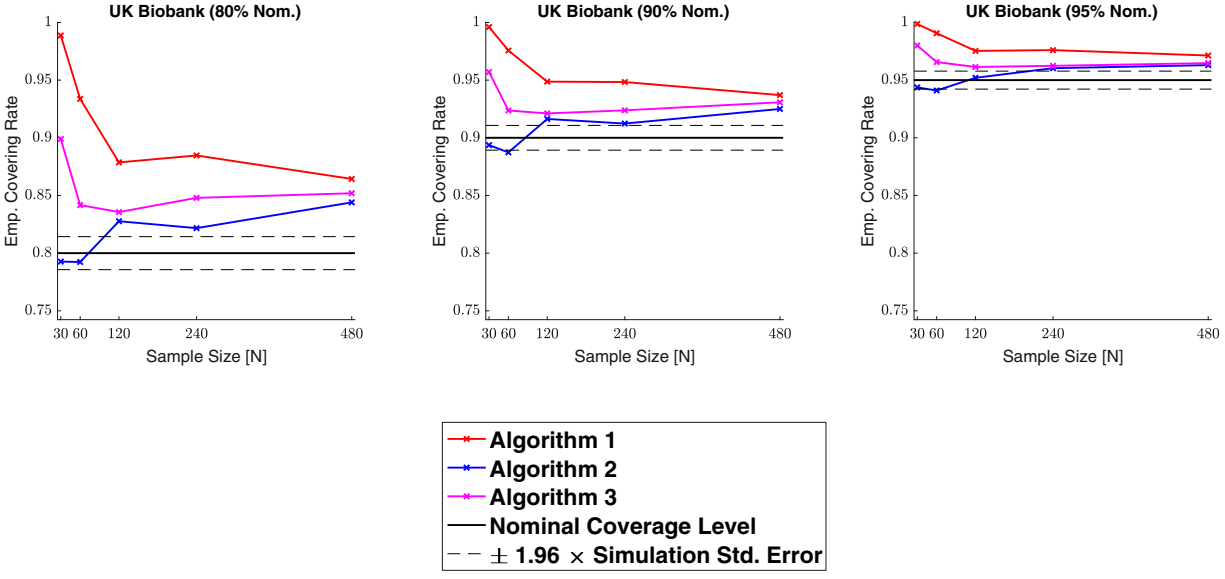


Fig. 11. Coverage results for the UK Biobank signal type, where the full standard deviation image was used as the standard deviation of the subject-level noise fields. Coverage results here were similar to the results for the multiple spheres signal type shown in Fig. 10. Once again, both Algorithms 2 and 3 performed well for large samples, with empirical coverage rates hovering above the nominal target, while results for Algorithm 1 came further above the nominal level. While for smaller samples the degree of over-coverage became greater for Algorithms 1 and 3, results for Algorithm 2 appear to slightly drop here.

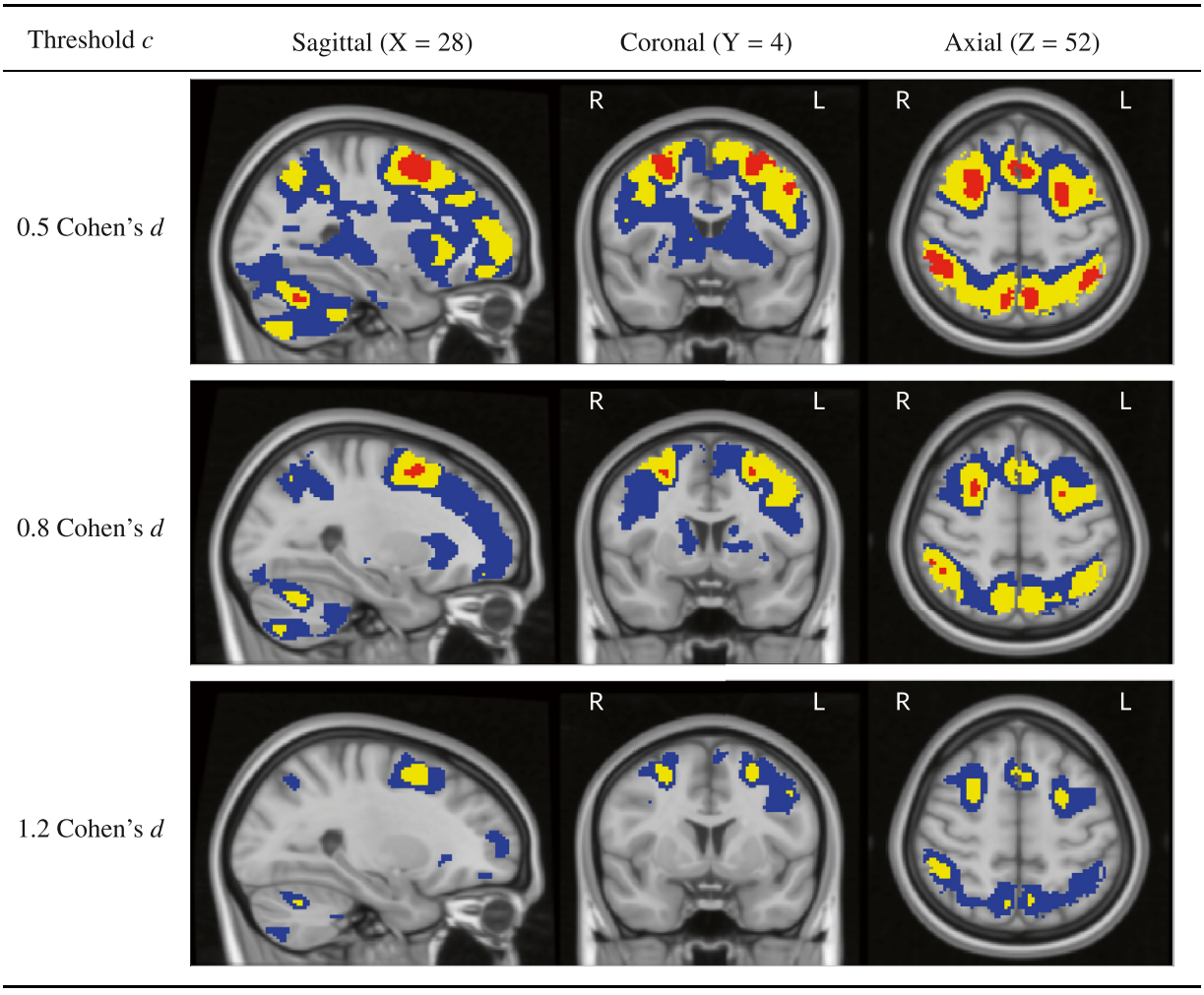


Fig. 12. Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 3 to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. The upper CS $\hat{\mathcal{A}}_c^+$ is displayed in red, and the lower CS $\hat{\mathcal{A}}_c^-$ in blue. Yellow voxels represent the point estimate set $\hat{\mathcal{A}}_c$, the best guess from the data of voxels that have surpassed the Cohen's d threshold. The red upper CS has localized regions in the frontal gyrus, paracingulate gyrus, angular gyrus, cerebellum and precuneus which we can assert with 95% confidence have attained (at least) a 0.5 Cohen's d effect size.

By increasing the threshold to $c = 0.8$ (Fig. 12, middle plot), there was a shrinking of both the blue lower CSs and red upper CSs. Therefore, while we can confidently declare a medium effect size in all of the brain areas identified above, the quantity of voxels within each region that we can proclaim to have a large effect size is considerably smaller. In the case of the right cerebellar hemisphere (left, sagittal slice) and insular cortex, the upper CS vanished completely, indicating that the method did not locate any voxels in these regions where we can assert a Cohen's d effect size greater than 0.8.

The results for the largest threshold assessed, $c = 1.2$ (Fig. 12, bottom plot) are particularly notable. Here, the yellow point estimate set contains a small but appreciable number of voxels, marking where the *observed* Cohen's d effect size was greater than 1.2. However, in this case there was *no* red upper CS (i.e. the upper CS was completely empty). Therefore, contrary to the inference one might come to based on the point estimate set alone, we can not state with confidence that *any* voxels have attained a very large effect size. Conversely, the large quantity of (grey background) voxels lying outside the blue lower CS in Fig. 12 imply an effect size *less* than 1.2 across the vast majority of the brain.

In Fig. 13, the red upper CSs computed with Algorithm 3 are compared with the thresholded t-statistic map (green-yellow voxels) obtained from applying a one-sample t-test group-analysis to the 80 subjects' contrast data, using a voxelwise FWE-corrected threshold of $p < 0.05$. This figure demonstrates the improved spatial specificity that can be provided with the CSs in comparison with the traditional approach. Specifically, while the thresholded statistic map contains one large cluster covering a sizeable portion of the parietal lobe across both brain hemispheres, the red upper CSs pinpoint precise areas in the precuneus and angular gyrus where a practically significant medium (or large) Cohen's d effect size can be inferred (Fig. 13, axial slices).

5. Discussion

5.1. Spatial inference on Cohen's d effect size

To fully appreciate the outcomes of a neuroimaging study, information about the *magnitude* (as well as presence) of effects must be reported at the end of an investigation. It is only with this knowledge that one can truly determine the practical relevance (and potential clinical importance) of any discoveries made during the analysis. In this work,

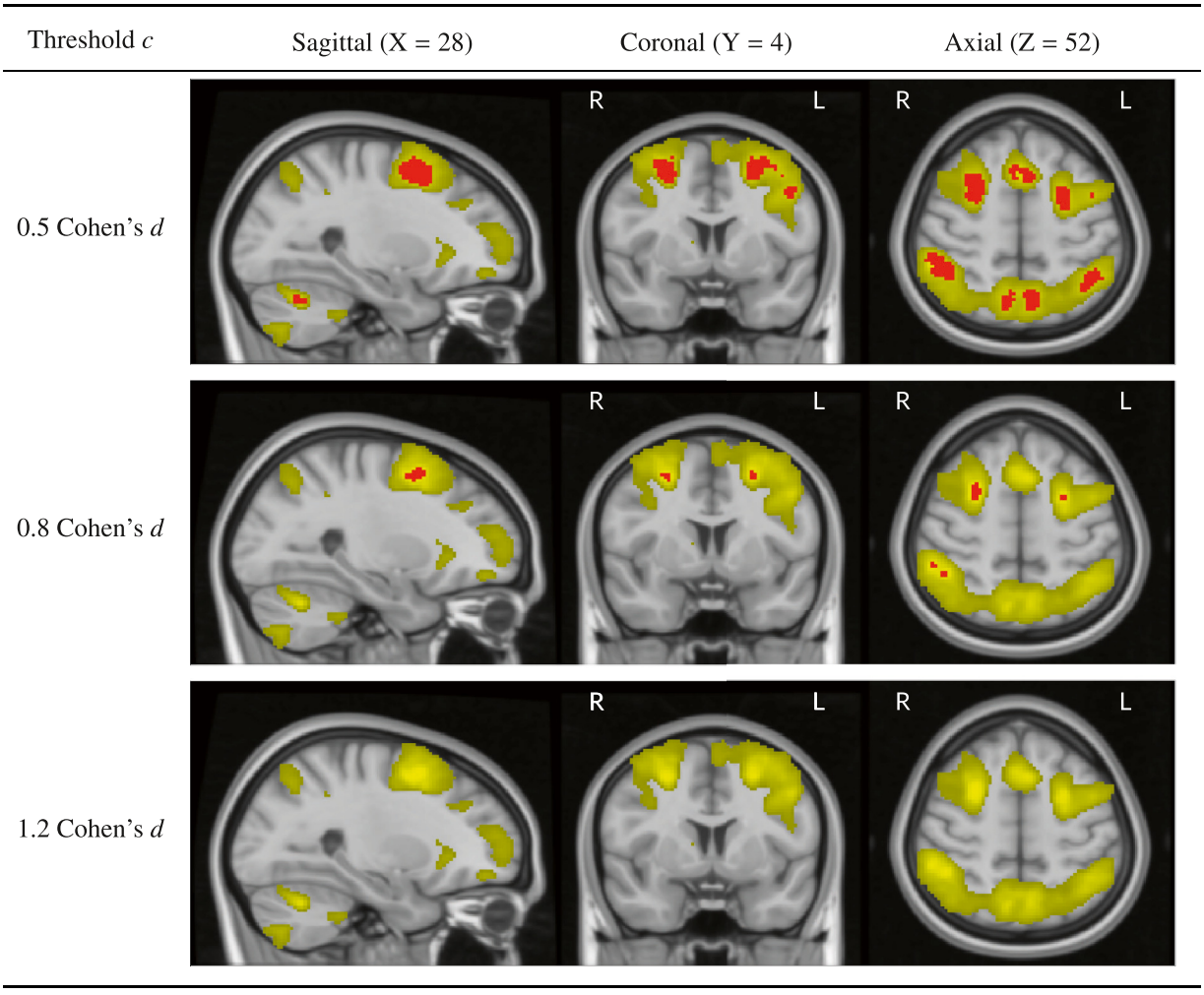


Fig. 13. Comparing the upper CSs (red voxels) computed with Algorithm 3 on the HCP working memory task data (same slice views as Fig. 12) with the thresholded t-statistic results obtained by applying a traditional group-level one-sample t-test, voxelwise $p < 0.05$ FWE correction (green-yellow voxels). While the thresholded statistic map contains a single cluster covering a sizable portion of the parietal lobe across both hemispheres (axial slices), the upper CSs have localized the precise areas of the precuneus and anglur gyrus where we can confidently declare a Cohen's d effect size of at least 0.5. This demonstrates how the CSs can provide improved spatial specificity in determining regions with practically significant activation.

we have presented three methods to create Confidence Sets for Cohen's d effect size maps, providing formal confidence statements on regions of the brain where the Cohen's d effect size has exceeded a specified activation threshold, alongside regions where the effect size has *not* surpassed this threshold. Both of these statements are made simultaneously across the entire brain, enabling researchers to pinpoint the precise regions where meaningful differences have occurred as well as identifying areas that have not responded to the task. This is in contrast to the traditional statistical approach, where a test statistic (e.g. a t-statistic) or a p-value can only quantify the compatibility between the observed data and what would be expected under the null-hypothesis of no activation, and no information is provided about the effect size when a finding is deemed to be statistically significant. Since the test statistic is confounded by the sample size of the study, it is not possible to implement a similar framework to obtain Confidence Sets for statistic images; as sample sizes increase test statistics also become arbitrarily large, until ultimately there is enough statistical power to declare even the smallest of effects as statistically significant. On the other hand, for larger samples we expect the red upper CSs to converge towards the population parameter $\mathcal{A}_c$ representing the set of voxels with a true population effect magnitude above a purposeful activation threshold c.

While in our analysis of the HCP working memory task data we primarily focused on activated regions localized by the red upper CSs, it is also important to note the added utility provided by the blue lower CSs that can not be obtained with standard statistical inferences. In particular, while the logic of statistical testing means that one can never prove the null-hypothesis to be true, the blue lower CSs provide inference on brain areas where effect sizes have *not* attained a sufficient activation threshold. Therefore, if the Cohen's d threshold is chosen appropriately, users could essentially accept the null of no (practically significant) activation for all voxels lying outside the blue CSs. Another method that facilitates for evidence in support of either the null- or an alternative hypothesis is Bayesian testing (Han and Park, 2018; Rouder et al., 2009). However, this procedure can not guarantee the same control of false-positives that has been demonstrated in terms of coverage for the CSs here, and requires that users set a log odds threshold (to define the strength of evidence needed to accept the null or alternative hypothesis) which is arguably less intuitive than the pre-determined confidence level needed for the CSs.

The use of CSs for inference on effect size may also help to alleviate issues associated with hypothesis testing for studies with lower statistical power. Specifically, for studies with small sample sizes it has been

reported that applying traditional inference procedures can lead to spurious or irreproducible results with considerably inflated observed effect sizes (Cremers et al., 2017; Poldrack et al., 2017). In regards to the latter, this is often caused by a form of selection bias known as the 'winners curse' (Button et al., 2013; Reddan et al., 2017), whereby voxels whose observed effect size has exceeded their expected performance are intrinsically more likely to be determined as statistically significant. This becomes a problem as magnitudes are commonly reported *only* for significant voxels, a practice that leads to positively biased effect estimates. Our analysis results from the Human Connectome Project dataset exemplify how the CSs can help to resolve this issue. In Fig. 12, the yellow point estimate 'best guess from the data' clusters identified a number of voxels with a Cohen's d effect size greater than 1.2 that were also included in the thresholded statistic map obtained from applying a one-sample t-test, voxelwise $p < 0.05$ FWE correction (Fig. 13). However, by synthesizing information about the effect magnitude as well as the *reliability* of the estimate, the CSs presented in Fig. 12 affirm that there are in fact *no* voxels that can be confidently declared to have an effect size larger than 1.2. On the contrary, only a handful of brain regions were contained in the red upper CS asserting a Cohen's d effect size exceeding 0.5 for the HCP working memory task data.

In our previous effort we described a method to obtain CSs for unstandardized percentage BOLD change maps, rather than the Cohen's d images that have served as our main focus here. The use of Cohen's d instead of %BOLD is likely to be advantageous due to complications associated with the BOLD effect. As discussed in our previous work, %BOLD effect sizes have been shown to modulate according to acquisition parameters such as the scanner field strength or MRI pulse sequence, and inhomogenieties in vascularity between different brain regions can cause further variation in the BOLD response. At a more rudimentary level, there are also difficulties involved in obtaining percentage BOLD change images. While all of the main fMRI software packages provide contrast of parameter estimate maps, each of the three most widely-used analysis packages (AFNI, FSL and SPM) scale the raw data differently; the parameter estimates are often given in arbitrary units which deviate between packages. Conversion to percentage BOLD change therefore requires a software-dependent normalization, where one must take into consideration how to appropriately scale the data, design matrix and analysis contrasts. While this can be cumbersome and prone to human error, conversion to Cohen's d is relatively simple. Due to the straightforward relationship between the Cohen's d effect size and the one-sample t-statistic ($\hat{d} = t/\sqrt{N}$), users can easily generate Cohen's d images from the unthresholded t-statistic maps created by all the main neuroimaging packages. For all of these reasons, the Cohen's d CS maps may be more suitable for comparison between studies.

In this work, we have used classifications of the Cohen's d effect size as initially suggested in Cohen (2013), describing 0.5 as a 'medium' effect, 0.8 as 'large', and 1.2 as 'very large'. While these benchmarks provide basic descriptors of effect size, in general we recommend that users take appropriate steps to contextualize what sort of magnitude constitutes a meaningful finding in their own study. In context, this means that for some task-fMRI studies brain regions with a Cohen's d effect size of 0.2 (classified as a 'small' effect by Cohen) may represent an important finding (see Fig. 2 in Noble et al., 2020 for the distributions of ground-truth effect sizes found across the brain for a range of common task-fMRI paradigms). Overall, users should factor in the aims of their investigation, the quality of the study and, if possible, the effect sizes reported in similar previous efforts before choosing a threshold. Obtaining the CSs for the Human Connectome Project contrast data in this work was computationally quick, with each analysis taking less than one minute for all three proposed algorithms. Therefore, one possible strategy is to evaluate a variety of different c's on pilot or historical data before fixing a value to use on a study of interest.

While we have developed the Cohen's d CSs for a one-sample model, the methods presented here may also be applied to the general linear model $\boldsymbol{Y}(s) = \boldsymbol{X}\boldsymbol{\beta}(s) + \epsilon(s)$, where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{\beta}(s)$ is the vector of unknown coefficients. In this setting, for a contrast vector $\boldsymbol{w}$, the quantity of interest would be the standardized contrast $\boldsymbol{w}^T\boldsymbol{\beta}(s)/\sigma(s)$ (instead of the Cohen's d effect size $\mu(s)/\sigma(s)$). The method would be carried out similarly, except, for example, the normalized standard deviation of the contrast estimate $\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}$ would need to be considered in the construction of the CSs (replacing the $1/\sqrt{N}$ term in the one-sample CS constructions for the three algorithms). It is important to note that the mathematical results underpinning this work in Telschow et al. (2020) have as yet only been provided for the one-sample model, which is why this model has been our primary focus here. Nevertheless, a further intuition on applying the method to the general linear model can be obtained from *BTSN*, where we developed the CSs in the general linear model setting for raw effect size images.

5.2. *Three algorithms for Cohen's* d *Confidence Sets*

In this work, we have theoretically motivated three algorithms for obtaining Cohen's d CSs. Our simulation results in Sections 4.1 and 4.2 have demonstrated differences in the coverage performance for each of these algorithms. Across all sets of simulation results, empirical coverage for Algorithm 1 came above the nominal level, with particularly severe over-coverage for 3D simulations carried out on large synthetic signals when small sample sizes were used (Figs. 8–10). The cause for such poor performance here is likely to be due to the variance term used to construct the CSs in Algorithm 1. Recalling the derivations in Section 2.2, the variance term $\sqrt{1 + \frac{\hat{d}^2(s)}{2}}$ used for Algorithm 1. was chosen as an estimator of the variance of the *limiting* Gaussian field $\mathcal{G}(s)$. Therefore, while we expect this term to correctly approximate the variance of the bootstrap approximating field asymptotically, our theory provides no indication about the accuracy of this term in small samples. The over-coverage seen in our simulation results suggests that this term overestimates the true variance of the approximating field when the sample size is low. While there was some improvement in Algorithm 1's results as N increased, even for the largest sample size we analyzed, $N = 480$, empirical coverage for the other two methods was consistently closer to the nominal target level.

Overall, Algorithm 2 and Algorithm 3 both performed well across our 2D and 3D simulations. For simulations using a 95% confidence level, the empirical coverage performance of these two methods was remarkably similar for $N \geq 120$ (in most cases, slightly above the nominal target). It is therefore difficult to conclude which method should be implemented in practice. For our 3D simulations (Figs. 8–11), Algorithm 2's empirical coverage results fell slightly closer to the nominal level in most cases. However, the results for Algorithm 3 were slightly more robust across moderate and large sample sizes, as observed in the UK Biobank and multiple spheres simulations (Figs. 10 and 11). When smaller sample sizes of $N \leq 60$ were used, Algorithm 3 consistently suffered from a higher degree of over-coverage. However, the performance of Algorithm 2 also appeared to be drop off in some cases, marginally for the UK Biobank simulation (Fig. 11), and considerably for the Ramp simulation (Fig. 6), where for $N = 30$ Algorithm 2's coverage results fell well below the nominal target level. Therefore, Algorithm 3 may still be preferable in small sample sizes to ensure the inference remains valid (in respect to obtaining at least a 95% coverage rate).

From a theoretical standpoint, the variance-stabilizing transformation approach used in Algorithm 3 assumes that the observations are Gaussian, while this is somewhat relaxed for Algorithm 2, where the bootstrap is applied to estimate the standard deviation directly from

the data. While this supports that Algorithm 2 may be preferable for non-Gaussian data, in our Human Connectome Project analyses the CSs maps obtained using both methods (Fig. 12 for Algorithm 3, Fig. D.2 for Algorithm 2) were virtually identical, indicating that both methods could be equally effective for fMRI data with sample sizes on the order of the HCP.

While our simulations have primarily focused on how well the three algorithms are able to maintain a targeted nominal coverage level, we also carried out further analyses to evaluate each method's capacity for identifying voxels with a true effect size greater than the threshold c (akin to the power of a statistical test). In Appendix E, we present the '$\hat{\mathcal{A}}_c^+$ sensitivity' of each of the three algorithms across our 2D and 3D simulations. For each simulation type, we computed the sensitivity of the upper CS $\hat{\mathcal{A}}_c^+$ as the proportion of voxels in the noise-free population cluster $\mathcal{A}_c$ that were identified in $\hat{\mathcal{A}}_c^+$. In other words, $\hat{\mathcal{A}}_c^+$ sensitivity is the percentage of voxels with a true Cohen's d effect size greater than c correctly identified within the upper CS $\hat{\mathcal{A}}_c^+$. Overall, the results for all three methods were similar here (Figs. E.1–E.6), although Algorithm 3 performed marginally better than the two remaining methods. In general, we found that more than 100 subjects were required to achieve a sensitivity above 10%, and that sensitivity could be exceedingly low when $N < 60$. These small-sample results are similar to corresponding sensitivity figures reported in Noble et al. (2020) for standard fMRI statistical methods, where the mean true positive rate was found to be less than 5% after a cluster-based FWER correction for a range of task-fMRI studies using a sample size of $N = 20$ (although note that this is not a like-for-like comparison, as the CSs control the inclusion error of individual voxels rather than clusters). While the CSs may not be very sensitive by this measure in small samples, we stress that the inference is still valid, in the sense that the CSs can still provide (at least) a 95% empirical coverage rate. Therefore, while a traditional statistical test with voxelwise FDR correction may be more suitable for detecting non-zero effects in this setting, the CSs could still find utility for delineating effect sizes that have surpassed a practically significant threshold and providing information about the spatial uncertainty of the thresholded clusters. We are pursuing further work on 'FDR-controlled' CSs, which we hope will lead to a more sensitive approach with the CSs in the future.

Although we have assessed the three algorithms on synthetic data, where the variance of the subject-level errors was either homo- or heterogeneous across space, a limitation of this work is that *only* Gaussian noise fields were considered for our simulations. While various non-Gaussian error fields were included in the simulations conducted in Sommerfeld et al. (2018), where the CSs were shown to be effective for inference on the sample mean effect size, further assessments on the methods proposed here using noise structures more comparable to actual fMRI data would be a valuable addition to any future work. As already discussed, the variance-stabilizing transformation utilized by Algorithm 3 makes an assumption of Gaussianity, so a future study may seek to verify whether this method's performance is adversely affected in a non-Gaussian setting. One approach that could be taken here is through empirical evaluations based on real data, where a large dataset is split in half to define a ground-truth and draw many random samples. We investigated such an evaluation with the UK Biobank, but found that even with 4000 subjects set aside we were unable to obtain a highly stable ground-truth set $\mathcal{A}_c$ required to accurately assess the coverage rate. As the UK Biobank expands towards making 100,000 subjects' fMRI data available, we look forward to revisiting this approach with yet larger sample sizes.

It is noticeable that the asymptotic coverage for all three procedures appeared to converge to above the nominal level in nearly all of our simulations. As well as this, the size of the over-coverage varied depending on the signal type and the confidence level (in all cases, the degree of over-coverage was higher for the smaller confidence level of $1 - \alpha = 0.80$ compared to the larger confidence level of $1 - \alpha = 0.95$). We do not believe this is due to changes in the signal type per se, but instead due to inaccuracies in the interpolation method used to assess if coverage was obtained (i.e. if $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$) caused by the resolution of the lattice, positively biasing results for signals with a longer boundary $\partial\mathcal{A}_c$. The assessment method evaluates whether coverage holds at a discrete set of sub-sampled points on $\partial\mathcal{A}_c$, but as the boundary length becomes longer, this set of discrete points becomes relatively less dense within the true, continuous boundary. Violations of the subset condition are therefore more likely to be missed for signals with a longer boundary, which could explain why, for example, the empirical coverage results for the large spherical effect were systematically higher than for the small spherical effect (Figs. 8 and 9), or why the coverage results for the UK Biobank signal type and a threshold of $c = 0.5$ (longer boundary) were systematically higher than the corresponding results obtained using a threshold of $c = 0.8$ (shorter boundary; Figs. C.15 and 11). This line of reasoning is also consistent with the greater levels of over-coverage seen for the lower confidence levels, as more violations of coverage should have occurred here, but this also meant there was a higher chance that violations could be missed. We discussed this issue in further detail in Section 5.2 of *BTSN*.

Data availability

We have used data from The Human Connectome Project and UK Biobank. All code used for the simulations and analysis of HCP data are available at: https://github.com/AlexBowring/Confidence_Sets_Manuscript. All Cohen's d Confidence Sets and statistical results maps obtained from our HCP analyses have been made available at the NeuroVault repository: https://neurovault.org/collections/9019/.

Credit authorship contribution statement

Alexander Bowring: Writing - original draft, Writing - review & editing, Methodology, Software, Formal analysis, Investigation, Visualization, Project administration. **Fabian J.E. Telschow:** Methodology, Software, Writing - review & editing. **Armin Schwartzman:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition. **Thomas E. Nichols:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition.

Acknowledgements

A.B. was supported by an Early Career Research Fellowship through the Nuffield Department of Population Health. T.E.N. was supported by the Wellcome Trust (100309/Z/12/Z). F.T., A.S. and T.E.N. were partially supported by NIH grant R01EB026859.

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

All authors would like to acknowledge the three anonymous reviewers, whose detailed feedback and insights (during the time of a global pandemic) helped to improve this manuscript.

Appendix A. Variance-stabilizing transformation of Cohen's d estimator

Theorem 1. *Assume the one-sample Gaussian model described in Section 2.2. For fixed N, let:*

$$a = \sqrt{\frac{N-1}{N-3}}, \qquad b = \sqrt{\frac{8N^2-17N+11}{(N-3)(4N-5)^2}},$$

and define

$$\alpha = \frac{1}{b}, \qquad \beta = \frac{a}{b}.$$

Then for $b^* = \sqrt{N}b, \alpha^* = \frac{\alpha}{\sqrt{N}}$, and $\beta^* = \sqrt{N}\beta$, define the transformation $\zeta : \mathbb{R} \to \mathbb{R}$ as:

$$\zeta\left(\hat{d}\right) = \sqrt{N}\left[\alpha^* arcsinh\left(\beta^*\hat{d}\right) - \alpha^* arcsinh\left(\beta^* d\left(1-\frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2N}b^{*2}\left(d\left(1-\frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}\right]. \tag{A.1}$$

Then the random variable $\zeta\left(\hat{d}\right)$ has, approximately, zero mean and unit variance.

Proof. We closely follow the workings given in the '*2. Noncentral t.*' section of Laubscher (1960). We have shown that $\sqrt{N}\hat{d}$ is distributed by a noncentral t-distribution with noncentrality parameter $\sqrt{N}\hat{d}$ and $N-1$ degrees of freedom. Defining:

$$C_N = \sqrt{\frac{N-1}{2}}\frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}, \tag{A.2}$$

where Γ is the gamma function, then the expectation and variance of $\sqrt{N}\hat{d}$ are:

$$\mathbb{E}\left[\sqrt{N}\hat{d}\right] = \sqrt{N}dC_N, \mathrm{Var}\left[\sqrt{N}\hat{d}\right] = \frac{N-1}{N-3} + Nd^2\left(\frac{N-1}{N-3} - C_N^2\right). \tag{A.4}$$

It is known that C_N is well-approximated by the polynomial

$$C_N \approx \left(1-\frac{3}{4N-5}\right)^{-1}. \tag{A.5}$$

Substituting into Eqs. (A.3) and (A.4), we deduce approximations of the expectation and variance of $\sqrt{N}\hat{d}$,

$$\mathbb{E}\left[\sqrt{N}\hat{d}\right] \approx \sqrt{N}d\left(1-\frac{3}{4N-5}\right)^{-1} = m_1, \tag{A.6}$$

$$\mathrm{Var}\left[\sqrt{N}\hat{d}\right] \approx \frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right) = m_2. \tag{A.7}$$

Now, noting that:

$$m_1^2 = Nd^2\frac{(4N-5)^2}{16(N-2)^2}, \tag{A.8}$$

then m_2 can be expressed in the form

$$m_2 = a^2 + b^2 m_1^2, \tag{A.9}$$

where $a = \sqrt{\frac{N-1}{N-3}}$, $b = \sqrt{\frac{8N^2-17N+11}{(N-3)(4N-5)^2}}$. Using the variance expression in (A.9) and applying Corollary 1 in Laubscher (1960), the approximate variance-stabilizing transformation of $\sqrt{N}\hat{d}$ is given by

$$\psi\left(\sqrt{N}\hat{d}\right) = \int_0^{\sqrt{N}\hat{d}}\left(a^2+b^2x^2\right)^{-\frac{1}{2}}dx = \alpha arcsinh\left(\beta\sqrt{N}\hat{d}\right), \tag{A.10}$$

where $\alpha = \frac{1}{b}$ and $\beta = \frac{a}{b}$. The quadratic Taylor approximation of $\psi\left(\sqrt{N}\hat{d}\right)$ about the point m_1 is given by

$$\psi\left(\sqrt{N}\hat{d}\right) \approx \psi(m_1) - \frac{1}{2}b^2 m_1 m_2^{-\frac{1}{2}}. \tag{A.11}$$

Therefore, the random variable:

$$Z\left(\sqrt{N}\hat{d}\right) = \psi\left(\sqrt{N}\hat{d}\right) - \psi(m_1) + \frac{1}{2}b^2 m_1 m_2^{-\frac{1}{2}} \tag{A.12}$$

will have, approximately, mean zero and unit variance. Substituting the precise expressions for ψ, m_1, and m_2 into (A.12) yields

$$Z\left(\sqrt{N}\hat{d}\right) = \alpha arcsinh\left(\beta\sqrt{N}\hat{d}\right) - \alpha arcsinh\left(\beta\sqrt{N}d\left(1-\frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2}b^2\left(\sqrt{N}d\left(1-\frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}. \tag{A.13}$$

At this point, we have established a variance-stabilizing transformation in terms of $\sqrt{N}\hat{d}$, when in practice, we require a transformation in terms of the Cohen's d estimator $\hat{d}$. This is possible by applying a change of variables to b, α and β. Defining $b^* = \sqrt{N}b$, $\alpha^* = \frac{1}{b^*} = \frac{1}{\sqrt{N}}\alpha$, and $\beta^* = \frac{b^*}{a} = \sqrt{N}\beta$, substituting into (A.13) obtains the desired transformation:

$$Z\left(\sqrt{N}\hat{d}\right) := \zeta\left(\hat{d}\right) = \sqrt{N}\left[\alpha^* arcsinh\left(\beta^*\hat{d}\right) - \alpha^* arcsinh\left(\beta^* d\left(1-\frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2N}b^{*2}\left(d\left(1-\frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}\right]. \tag{A.14}$$

□

Appendix B. UK Biobank Cohen's d image histogram

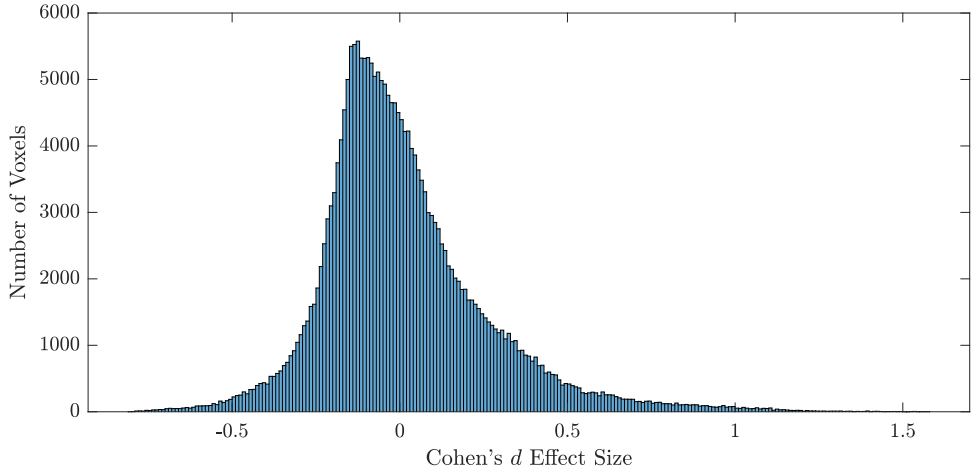


Fig. B1. Histogram showing the distribution of effect sizes in the UK Biobank Cohen's d field used for the final 3D simulation, as shown in the bottom row of Fig. 5.

Appendix C. Supplementary simulation results

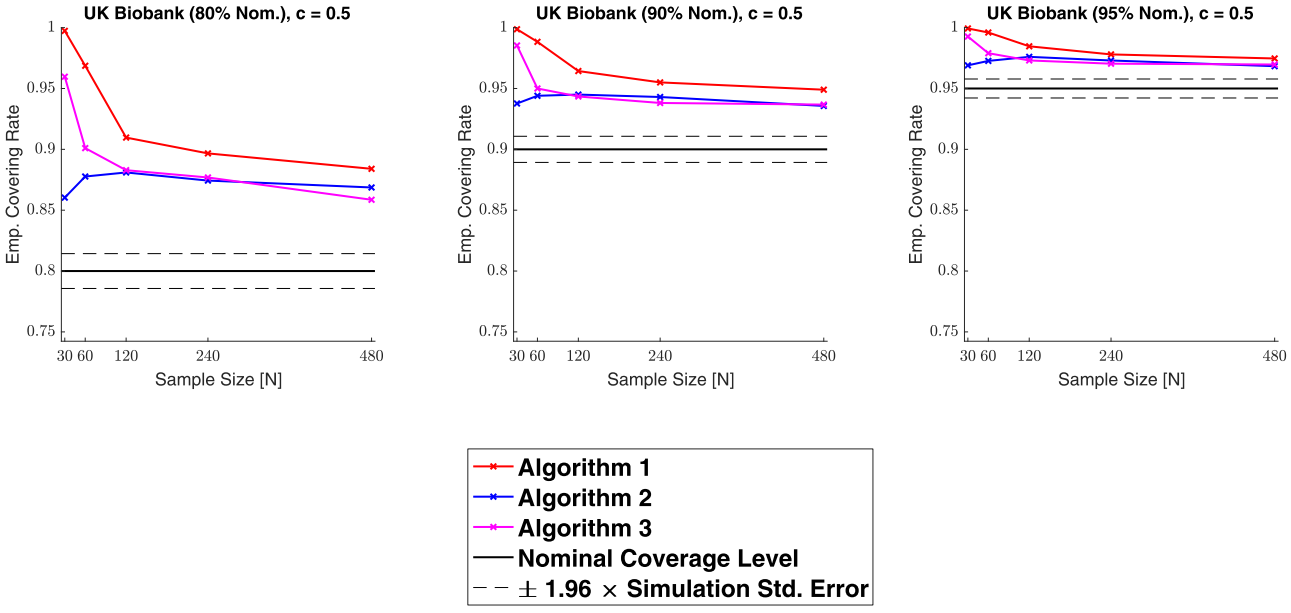


Fig. C1. Coverage results for the UK Biobank signal type simulation with a Cohen's d threshold of $c = 0.5$ (instead of the $c = 0.8$ threshold used for the simulation results presented in Fig. 11). For this smaller threshold, we observed valid, over-coverage for all three methods across all sample sizes, and on-the-whole the CSs performed well. In comparison to the results obtained for the larger threshold of $c = 0.8$ (Fig. 11), it is notable that there is a slightly higher degree of over-coverage across all of the results here. We believe this may be in part due to inaccuracies in the interpolation method used to assess the simulations results, rather than inaccuracies in the method itself; as the boundary length $\partial\mathcal{A}_c$ is longer for the smaller threshold used here, it is more likely that violations of coverage were missed (due to the fact coverage is assessed at only a discrete set of lattice points along $\partial\mathcal{A}_c$), inducing a positive bias in the results. We discuss this issue in more depth at the end of Section 5.2.

Appendix D. Supplementary Human Connectome Project results

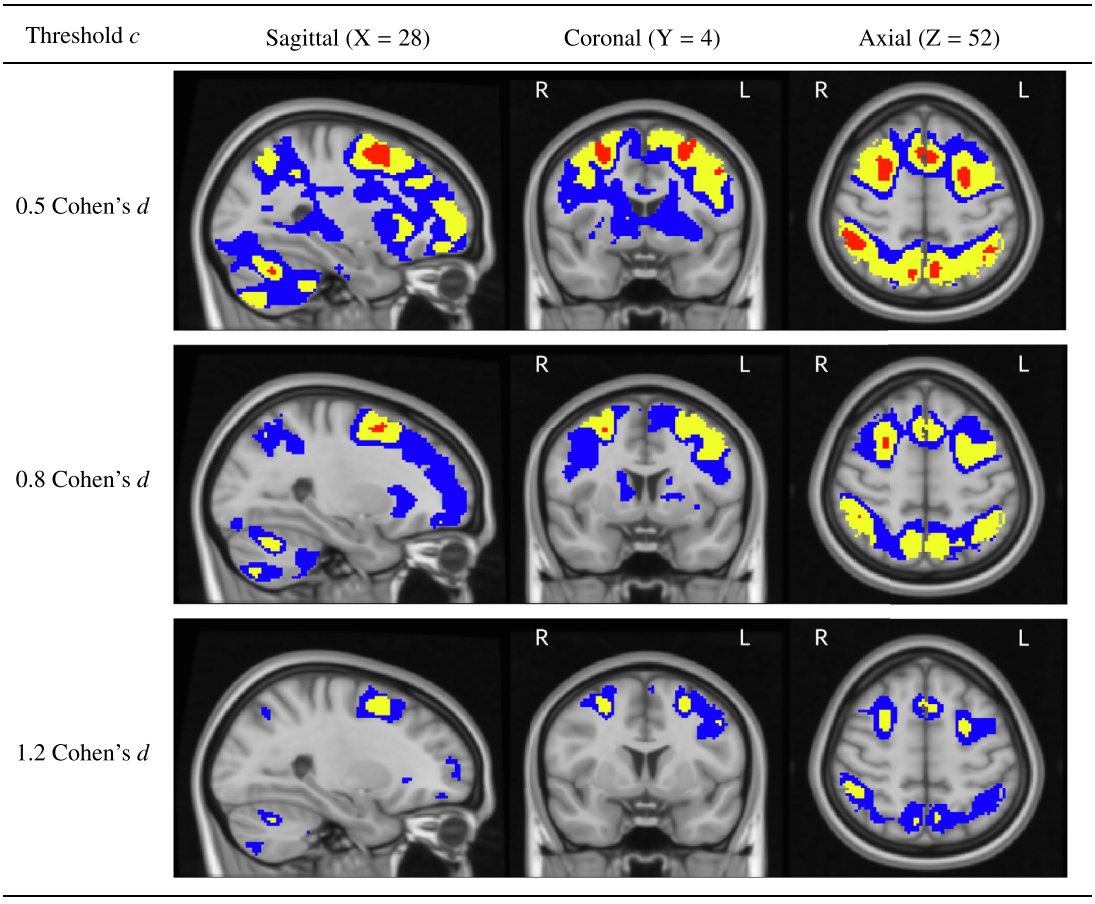


Fig. D1. S.1:Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 1 to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. Comparing with Fig. 12 and Fig. D.2, the CSs presented here are slightly more conservative than the corresponding CSs obtained with Algorithms 2 and 3 (in the sense that the red upper CSs here are smaller, and blue lower CSs are larger). This is consistent with the simulation results obtained in Sections 4.1 and 4.2, where the empirical coverage for Algorithm 1 was consistently larger than the other two methods.

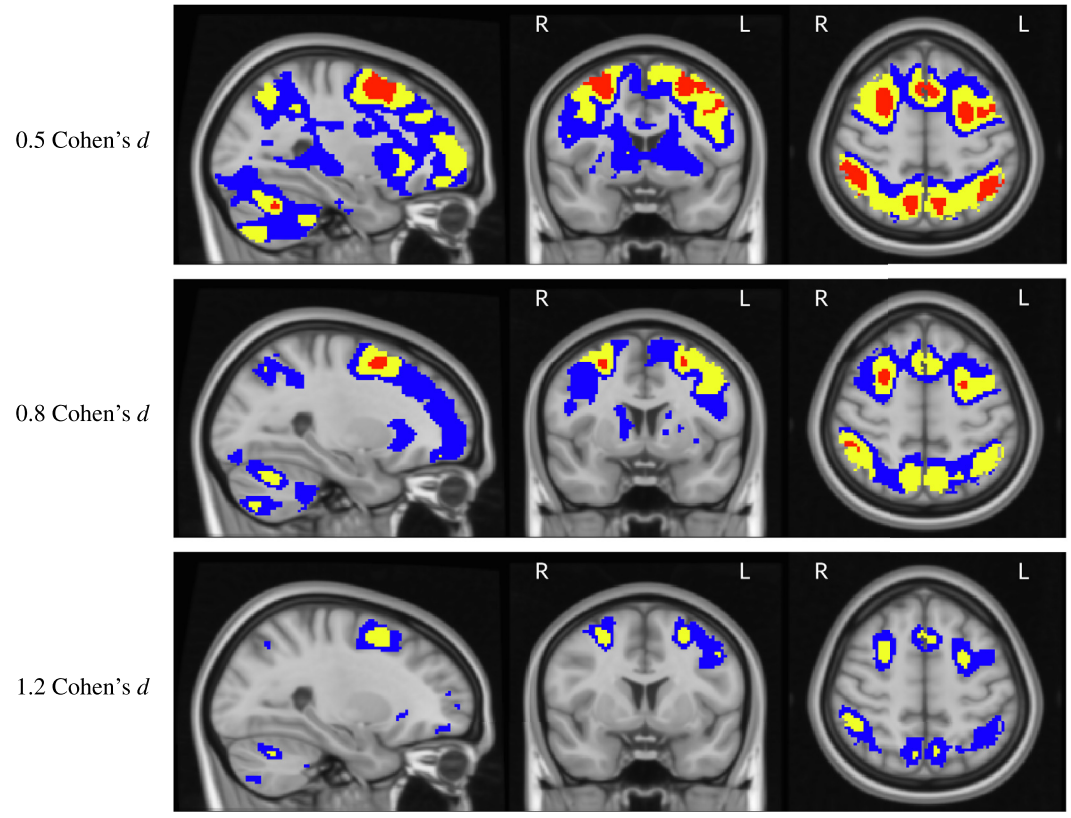


Fig. D2. Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 2 to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. Comparing with Fig. 12, the upper and lower CSs presented here are almost identical to the corresponding CSs obtained with Algorithm 3.

Appendix E. $\hat{\mathcal{A}}_c^+$ Sensitivity

Here we provide an indication of the sensitivity of the upper Confidence Sets $\hat{\mathcal{A}}_c^+$ obtained using the three algorithms proposed in Section 2.6 across our simulations. For each of the 2D and 3D simulation results presented in Sections 4.1 and 4.2, in Figs. E1, E2, E3, E4, E5 and E6 below we show the proportion of voxels in the noise-free population cluster $\mathcal{A}_c$ that were included in the upper CS $\hat{\mathcal{A}}_c^+$ across all toy-runs of the respective signal-plus-noise model whenever coverage was obtained (i.e. trials for which $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$). In other words, this is the average percentage of voxels with a true Cohen's d effect size greater than c that were correctly identified within the upper CS $\hat{\mathcal{A}}_c^+$. In this sense, these results are similar to the true positive rate used to indicate the sensitivity of a statistical test.

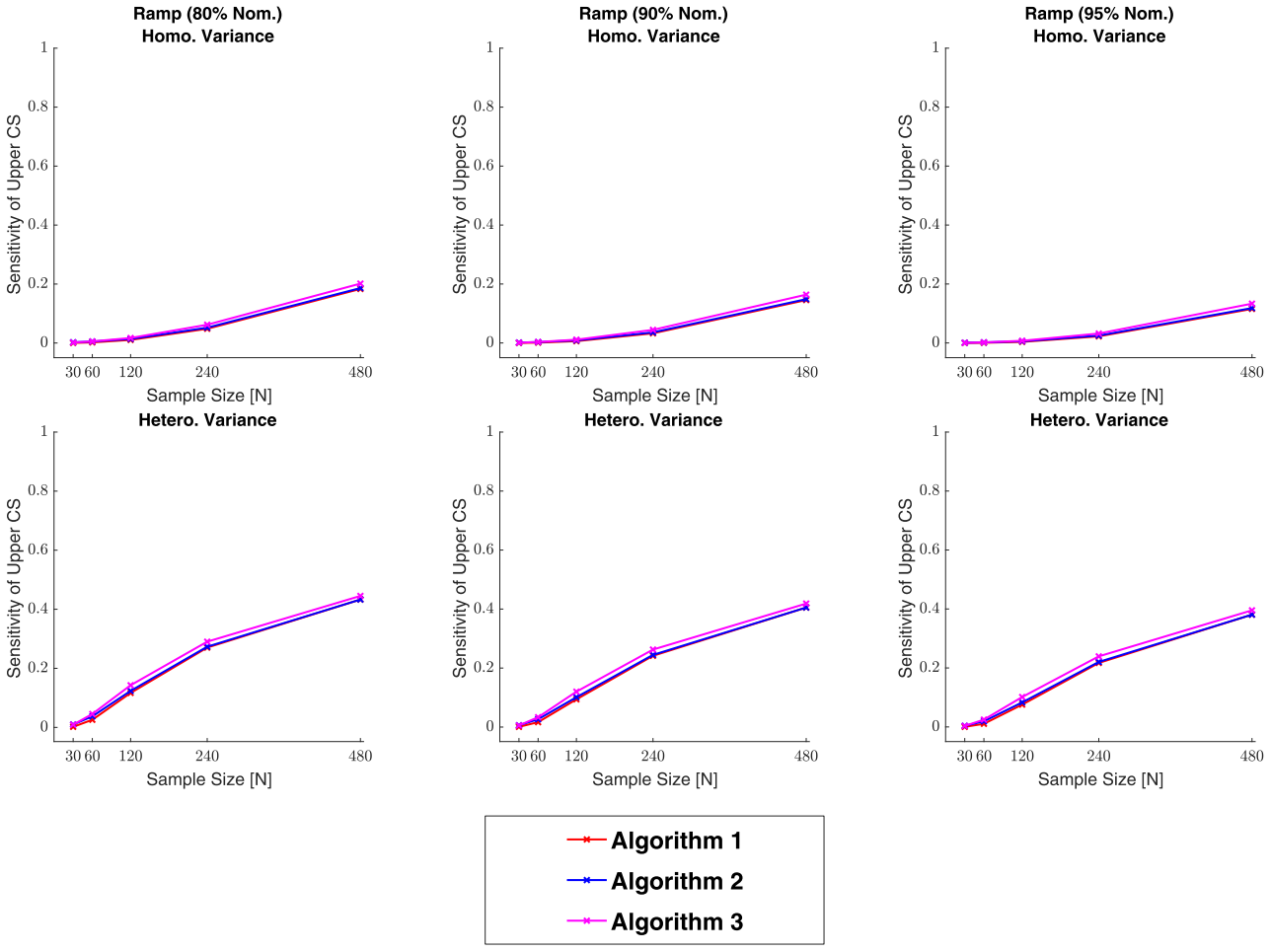


Fig. E1. Sensitivity results for the ramp signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures.

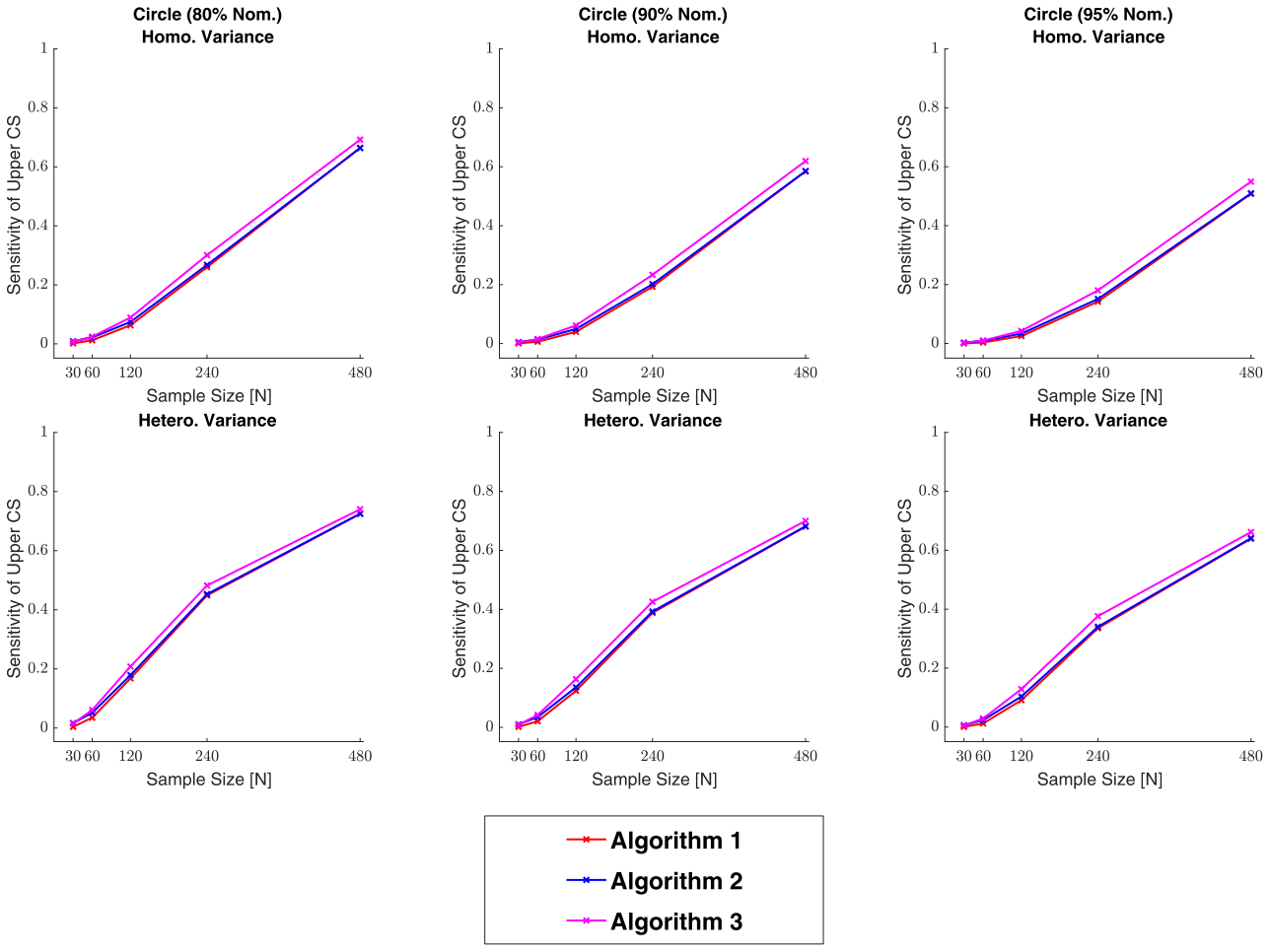


Fig. E2. Sensitivity results for the circular signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures.

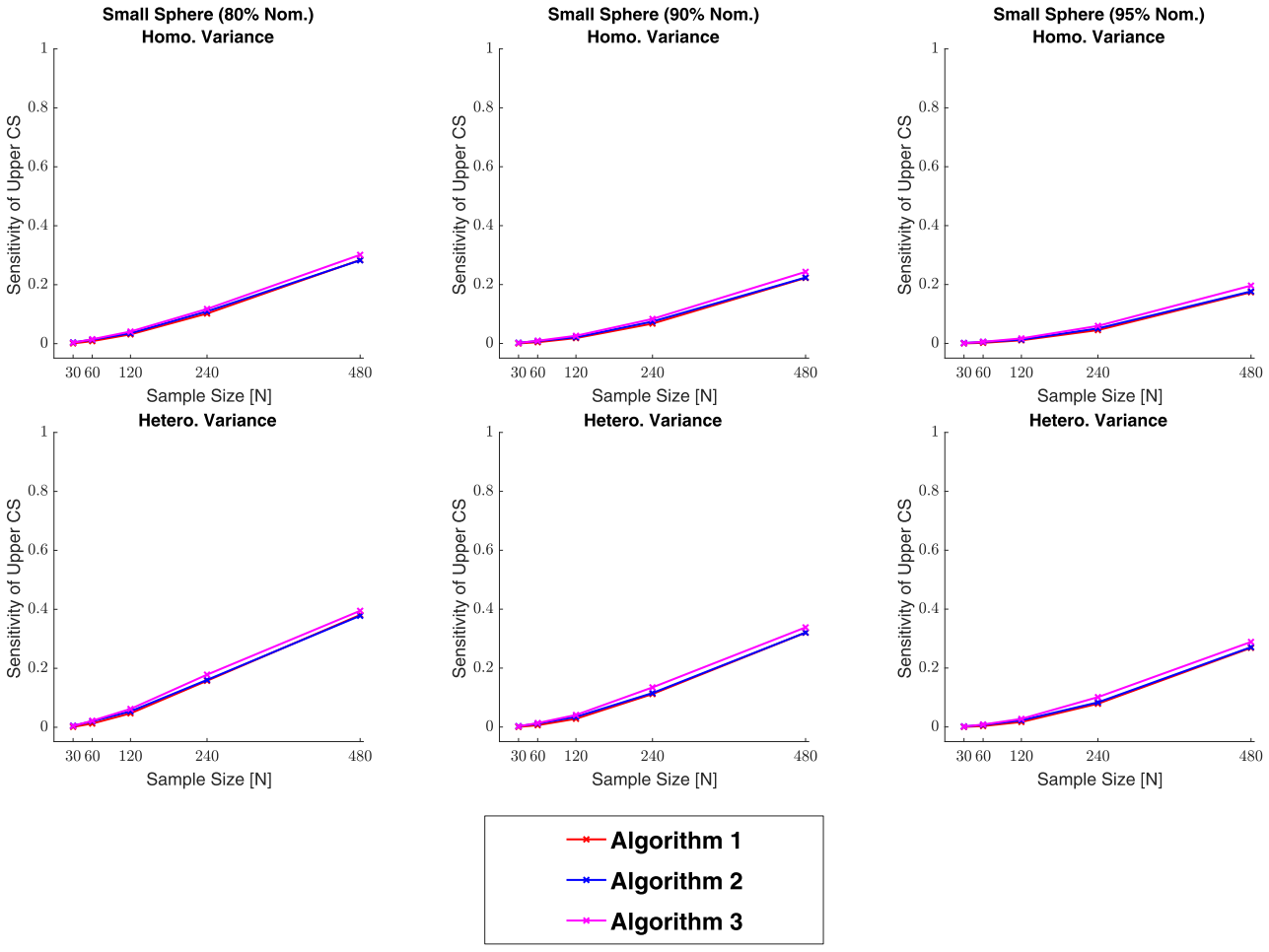


Fig. E3. Sensitivity results for the small sphere signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures.

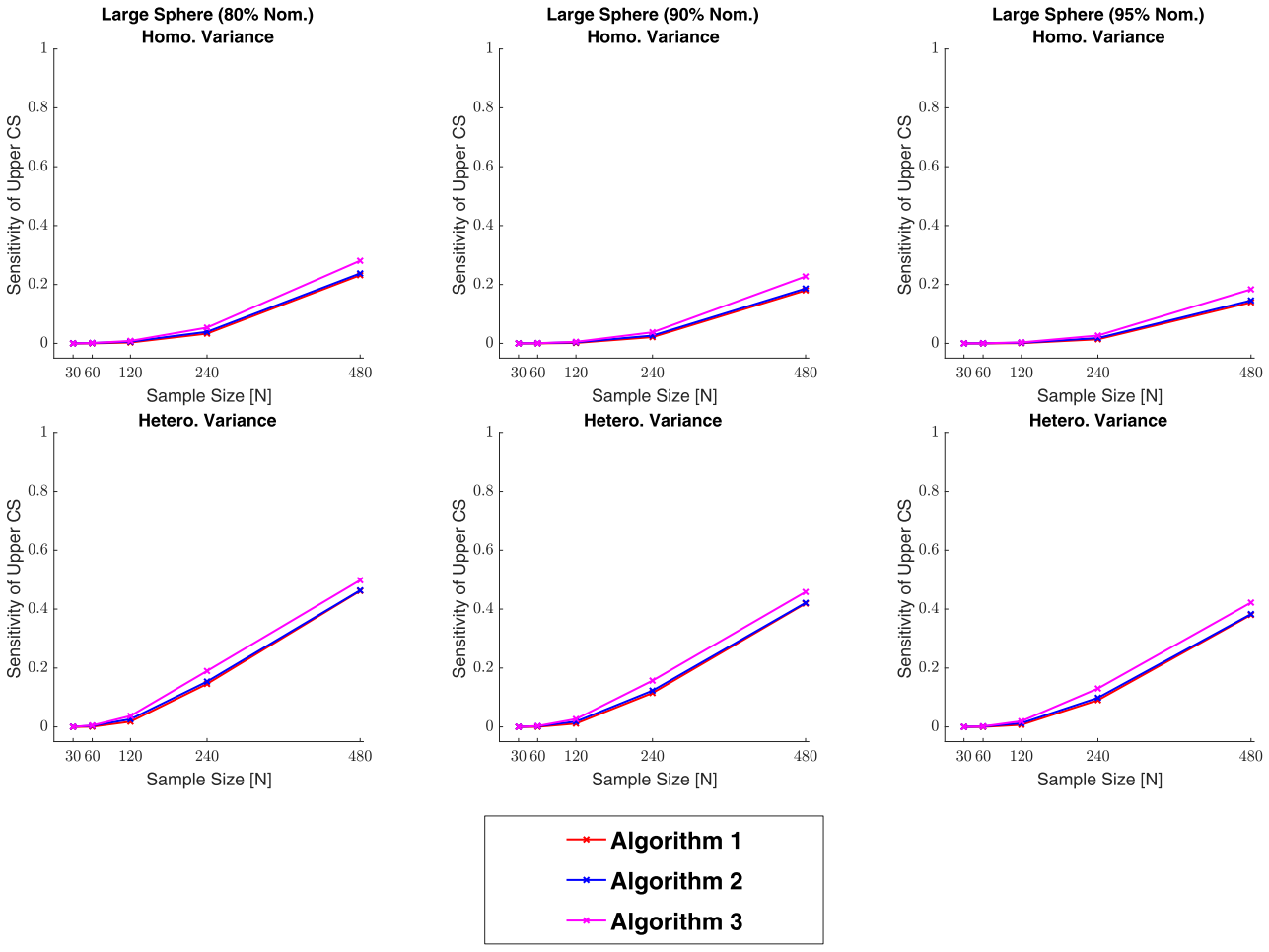


Fig. E4. Sensitivity results for the large sphere signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures.

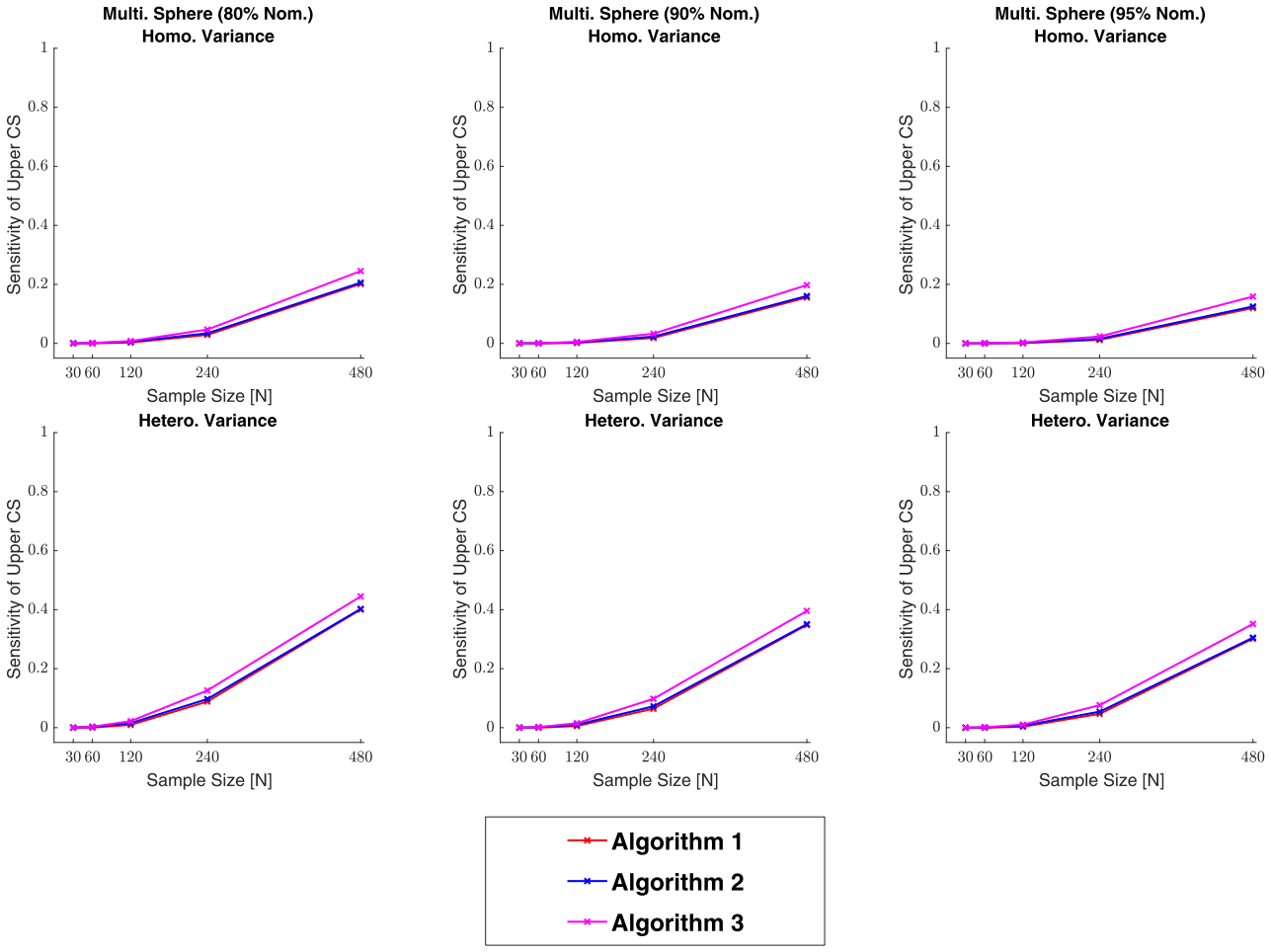


Fig. E5. Sensitivity results for the multiple spheres signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures.

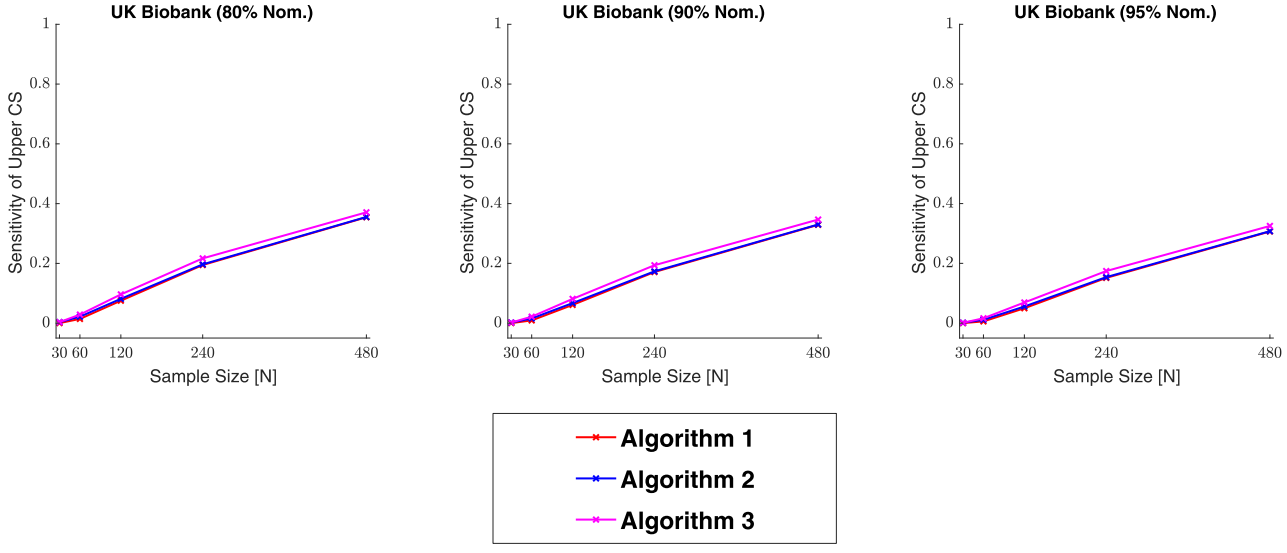


Fig. E6. Sensitivity results for the UK Biobank signal, where the UK Biobank full standard deviation image was used as the standard deviation of the subject-level Gaussian noise fields.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. NeuroImage 166, 400–424.

Bowring, A., Telschow, F., Schwartzman, A., Nichols, T.E., 2019. Spatial confidence sets for raw effect size images. NeuroImage 203, 116187. doi:10.1016/j.neuroimage.2019.116187.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14 (5), 365–376.

Cacioppo, J.T., Cacioppo, S., Gonzaga, G.C., Ogburn, E.L., VanderWeele, T.J., 2013. Marital satisfaction and break-ups differ across on-line and off-line meeting venues. Proc. Natl. Acad. Sci. U.S.A. 110 (25), 10135–10140.

Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. NeuroImage 63 (1), 289–300.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Routledge.

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

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., WU-Minn HCP Consortium, 2013. The minimal preprocessing pipelines for the human connectome project. NeuroImage 80, 105–124.

Gonzalez-Castillo, J., Saad, Z.S., Handwerker, D.A., Inati, S.J., Brenowitz, N., Bandettini, P.A., 2012. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proc. Natl. Acad. Sci. U.S.A. 109 (14), 5487–5492.

Han, H., Park, J., 2018. Using SPM 12s second-level Bayesian inference procedure for fMRI analysis: practical guidelines for end users. Front. Neuroinform. 12, 1. doi:10.3389/fninf.2018.00001.

Hariri, A.R., Tessitore, A., Mattay, V.S., Fera, F., Weinberger, D.R., 2002. The amygdala response to emotional stimuli: a comparison of faces and scenes. NeuroImage 17 (1), 317–323.

Laubscher, N.F., 1960. Normalizing the noncentral *t* and *F* distributions. The Annals of Mathematical Statistics 31 (4), 1105–1112.

Meehl, P.E., 1967. Theory-testing in psychology and physics: a methodological paradox. Philos. Sci. 34 (2), 103–115.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536.

Noble, S., Scheinost, D., Constable, R.T., 2020. Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. NeuroImage 209, 116468. doi:10.1016/j.neuroimage.2019.116468.

Nuzzo, R., 2013. In: Nature (Ed.), Online daters do better in the marriage stakes. http://www.doi.org/10.1038/nature.2013.13120. Accessed: 2020-04-04.

Nuzzo, R., 2014. Scientific method: statistical errors. Nature 506 (7487), 150–152.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18 (2), 115–126.

Reddan, M.C., Lindquist, M.A., Wager, T.D., 2017. Effect size estimation in neuroimaging. JAMA Psychiatry 74 (3), 207.

Rosenfeld, M.J., Thomas, R.J., Hausen, S., 2019. Disintermediating your friends: how online dating in the united states displaces other ways of meeting. Proc. Natl. Acad. Sci. U.S.A. 116 (36), 17753–17758.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonom. Bull. Rev. 16 (2), 225–237. doi:10.3758/PBR.16.2.225.

Rozeboom, W.W., 1960. The fallacy of the null-hypothesis significance test. Psychol. Bull. 57 (5), 416–428.

Sommerfeld, M., Sain, S., Schwartzman, A., 2018. Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. J. Am. Stat. Assoc. 113 (523), 1327–1340.

Telschow, F. J. E., Davenport, S., Schwartzman, A., 2020. Functional delta residuals and applications to functional effect sizes. 2005.10041.

Van Essen, D.C., Glasser, M.F., 2016. The human connectome project: progress and prospects. Cerebrum 2016.

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