Handling Multiplicity in Neuroimaging through Bayesian Lenses with Hierarchical Modeling

Gang Chen*[a], Yaqiong Xiao[b], Paul A. Taylor[a], Tracy Riggins[b], Fengji Geng[b], Elizabeth Redcay[b], and Robert W. Cox[a]

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, USA
[b]Department of Psychology, University of Maryland, USA

Abstract

In neuroimaging, the multiplicity issue may sneak into data analysis through several channels, affecting expected false positive rates (FPRs; type I errors) in diverse ways. One widely recognized aspect of multiplicity, multiple testing, occurs when the investigator fits a separate model for each voxel in the brain. However, multiplicity also occurs when the investigator conducts multiple comparisons within a model, tests two tails of a t-test separately when prior information is unavailable about the directionality, and branches in the analytic pipelines. The current practice of handling multiple testing through controlling the overall FPR in neuroimaging under the null hypothesis significance testing (NHST) paradigm excessively penalizes the statistical power with inflated type II errors. More fundamentally, the adoption of dichotomous decisions through sharp thresholding under NHST may not be appropriate when the null hypothesis itself is not pragmatically relevant because the effect of interest takes a continuum instead of discrete values and is not expected to be null in most brain regions. When the noise inundates the signal, two different types of error are more relevant than the concept of FPR: incorrect sign (type S) and incorrect magnitude (type M).

In light of these considerations, we introduce a different strategy using Bayesian hierarchical modeling (BHM) to achieve two goals: 1) improving modeling efficiency via one integrative (instead of many separate) model and dissolving the multiple testing issue, and 2) turning the focus of conventional NHST on FPR into quality control by calibrating type S errors while maintaining a reasonable level of inference efficiency. The performance and validity of this approach are demonstrated through an application at the region of interest (ROI) level, with all the regions on an equal footing: unlike the current approaches under NHST, small regions are not disadvantaged simply because of their physical size. In addition, compared to the massively univariate approach, BHM may simultaneously achieve increased spatial specificity and inference efficiency. The benefits of BHM are illustrated in model performance and quality checking using an experimental dataset. In addition, BHM offers an alternative, confirmatory, or complementary approach to the conventional whole brain analysis under NHST, and promotes results reporting in totality and transparency. The methodology also avoids sharp and arbitrary thresholding in the p-value funnel to which the multidimensional data are reduced. The modeling approach with its auxiliary tools will be available as part of the AFNI suite for general use.

Introduction

The typical neuroimaging data analysis at the whole brain level starts with a preprocessing pipeline, and then the preprocessed data are fed into a voxel-wise time series regression model for each subject. An effect estimate

*Corresponding author. E-mail address: gangchen@mail.nih.gov is then obtained at each voxel as a regression coefficient that is, for example, associated with a task/condition or a contrast between two effects or a linear combination among multiple effects. Such effect estimates from individual subjects are next incorporated into a population model for generalization, which can be parametric (e.g., Student's t-test, AN(C)OVA, univariate or multivariate GLM, linear mixed-effects (LME) or Bayesian model) or nonparametric (e.g., permutations, bootstrapping, rank-based testing). In either case, this generally involves one or more statistical tests at each spatial element separately.

As in many scientific fields, the typical neuroimaging analysis has traditionally been conducted under the framework of null hypothesis significance testing (NHST). As a consequence, a big challenge when presenting the population results is to properly handle the multiplicity issue resulting from the tens of thousands of simultaneous inferences, but this undertaking is met with various subtleties and pitfalls due to the complexities involved: the number of voxels in the brain (or a restricting mask) or the number of nodes on surface, spatial heterogeneity, violation of distributional assumptions, etc. The focus of the present work will be on developing an efficient approach from Bayesian perspective to address the multiplicity issue as well as some of the pitfalls associated with NHST. We first describe the multiplicity issue and how it directly results from the NHST paradigm and inefficient modeling, and then translate many of the standard analysis features to the proposed Bayesian framework.

Multiplicity in neuroimaging

In statistics, multiplicity is more often referred to as multiple comparisons or multiple testing problem when more than one statistical inference is made simultaneously. In general, we can classify four types of multiplicity issues that commonly occur in neuroimaging data analysis.

A) *Multiple testing.* The first and major multiplicity arises when the same design (or model) matrix is applied multiple times to different values of the response or outcome variable, such as the effect estimates at the voxels within the brain. As the conventional voxel-wise neuroimaging data analysis is performed with a massively univariate approach, there are as many models as the number of voxels, which is the source of the major multiplicity issue: multiple testing. Those models can be, for instance, Student's t-tests, AN(C)OVA, univariate or multivariate GLM, LME or Bayesian model. Regardless of the specific model, all the voxels share the same design matrix, but have different response variable values on the left-hand side of the equation. With human brain size on the order of 10^6 mm^3, the number of voxels may range from 20,000 to 150,000 depending on the voxel dimensions. Each extra voxel adds an extra model and leads to incrementally mounting odds of pure chance or "statistically significant outcomes," presenting the challenge to account for the occurrence of mounting family-wise error (FWE), while effectively holding the overall false positive rate (FPR) at a nominal level (e.g., 0.05). In the same vein, surface-based analysis is performed with 30,000 to 50,000 nodes (Saad et al., 2004), sharing a similar multiple testing issue with its volume-based counterpart. Sometimes the investigator performs analyses at smaller number of regions of interest (ROIs), perhaps of order 100, but even here adjustment is still required for the multiple testing issue (though it is often not made).

B) *Double sidedness.* Another occurrence of multiplicity is the widespread adoption of two separate one-sided (or one-tailed) tests in neuroimaging. For instance, the comparison between the two conditions of "easy" and "difficult" are usually analyzed twice for the whole brain: one showing whether the easy effect is higher than difficult, and the other for the possibility of the difficult effect being higher than easy. One-sided testing for one direction would be justified if prior knowledge is available regarding the sign of the test for a particular brain region. When no prior information is available for all regions in the brain, one cannot simply finesse two separate one-sided tests in place of one two-sided test, and a double sidedness practice warrants a Bonferroni correction because the two tails are independent with respect to each other (and each one-sided test is more liberal than a two-sided test at the same significance level). However, simultaneously testing both tails in tandem for whole brain analysis without correction is widely used without clear justification, and this forms a source of multiplicity issue that needs proper accounting.

C) *Multiple comparisons.* It rarely occurs that only one statistical test is carried out in a specific neuroimaging study, such as a single one-sample t-test. Therefore, a third source of multiplicity is directly related to the popular term, multiple comparisons. For example, an investigator that designs an emotion experiment with three conditions (easy, difficult, and moderate) may perform several separate tests: comparing each of the three conditions to baseline, making three pairwise comparisons, or testing a linear combination of the three conditions (such as the average of easy and difficult versus moderate). However, neuroimaging publications seldom consider corrections for such separate tests.

D) *Multiple paths.* The fourth multiplicity issue to affect outcome interpretation arises from the number of potential preprocessing, data dredging and analytical pipelines. For instance, all common steps have a choice of procedures: outlier handling (despiking, censoring), slice timing correction (yes/no, various interpolations), head motion correction (different interpolations), different alignment methods from EPI to anatomical data plus upsampling (1 to 4 mm), different alignment methods to different standard spaces (Talairach and MNI variants), spatial smoothing (3 to 10 mm), data scaling (voxel-wise, global or grand mean), confounding effects (slow drift modeling with polynomials, high pass filtering, head motion parameters), hemodynamic response modeling (different presumed functions and multiple basis functions), serial correlation modeling (whole brain, tissue-based, voxel-wise AR or ARMA), and population modeling (univariate or multivariate GLM, treating sex as a covariate of no interest (thus no interactions with other variables) or as a typical factor (plus potential interactions with other variables)). Each choice represents a "branching point" that could have a quantitative change to the final effect estimate and inference. Conservatively assuming three options at each step here would yield totally $3^{10} = 59,049$ possible paths, commonly referred to as researcher degrees of freedom (Simmons et al., 2011). The impact of the choice at each individual step for this abbreviated list might be negligible, moderate, or substantial; for example, the estimate for spatial correlation of the noise could be sensitive to the voxel size to which the original data were upsampled (Mueller et al., 2017; Cox and Taylor, 2017), which may lead to different cluster thresholds and poor control to the intended FPR in correcting for multiplicity. Therefore, the cumulative effect across all these hierarchical branching points could be a large divergence between any two paths for the final results. A multiverse analysis (Steegen et al., 2016) has been suggested for such situations of having a "garden of forking paths" (Gelman and Loken, 2013), but this seems highly impractical for neuroimaging data. Even when one specific analytical path is chosen by the investigator, it remains possible to invoke potential or implicit multiplicity in the sense that the details of the analytical steps such as data sanitation are conditional on the data (Gelman and Loken, 2013). The final interpretation of significance typically ignores the number of choices or the potential branchings that may affect the final outcome, even though it would be more preferable to have the statistical significance independent of these preprocessing steps.

The challenges of dealing with multiple testing at the voxel or node level have been recognized within the neuroimaging community almost as long as the history of FMRI. Substantial efforts have been devoted to ensuring that the actual type I error (or FPR) matches its nominal requested level under NHST. Due to the presence of spatial non-independence of noise, the traditional approach to countering multiple testing through Bonferroni correction is highly conservative, so the conventional correction efforts have been channeled into two main categories, 1) controlling for FWE, so that the overall FPR at the cluster or whole brain level is approximately at the nominal value, and 2) controlling for false discovery rate (FDR), which harnesses the expected proportion of identified items or discoveries that are incorrectly labeled (Benjamini and Hochberg, 1995). FDR can be used to handle a needle-in-haystack problem where a small number of effects existing among a sea of zero effects in, for example, bioinformatics. However, FDR is usually quite conservative for typical neuroimaging data and thus is not widely adopted. Therefore, we do not discuss it hereafter in the current context.

Typical FWE correction methods for multiple testing include Monte Carlo simulations (Forman et al., 1995), random field theory (Worsley et al., 1992), and permutation testing (Nichols and Holmes, 2001; Smith and Nichols, 2009). Regardless of the specific FWE correction methodology, the common theme is to use the spatial extent, either explicitly or implicitly, as a pivotal factor. One recent study suggested that the nonparametric methods seem to have achieved a more uniformly accurate controllability for FWE than their parametric counterparts

(Eklund et al., 2016), even though parametric methods may exhibit more flexibility in modeling capability (and some parametric methods can show reasonable FPR controllability; Cox et al., 2017). Because of this recent close examination (Eklund et al., 2016) on the practical difficulties of parametric approaches herein controlling FWE, there is currently a rule of thumb that demands any parametric correction be based on a voxel-wise p-value threshold at 0.001 or less. Such a narrow modeling choice with a harsh cutoff could be highly limiting, depending on several parameters such as trial duration (event-related versus block design), and would definitely make small regions even more difficult to pass through the NHST filtering system. In other words, the leverage on spatial extent with a Procrustean criterion undoubtedly incurs a collateral damage: small regions (e.g., amygdala) or subregions within a brain area are inherently placed in a disadvantageous position; that is, to be able to surpass the threshold bar, small regions would have to reach a much higher signal strength to survive a uniform criterion at the cluster threshold or whole brain level.

The concept of using contiguous spatial extent as a leveraging mechanism to control for multiplicity can be problematic from another perspective. For example, suppose that two anatomically separate regions are spatially distant and the statistical evidence (as well as signal strength) for each of their effects is not strong enough to pass the cluster correction threshold individually. However, if another two anatomically regions that have exactly the same statistical evidence (as well as signal strength) are adjacent, their spatial contiguity could elevate their combined volume to the survival of correction for FWE. Trade-offs are inherently involved in these final interpretations. One may argue that the sacrifice in statistical power under NHST is worth the cost in achieving the overall controllability of type I error, but it may be unnecessarily over-penalizing to stick to such an inflexible criterion rather than utilizing the neurological context or prior knowledge, as discussed below.

To summarize the debate surrounding cluster inferences, multiplicity is directly associated with the concept of false positives or type I errors under NHST, and the typical control for FWE at a preset threshold (e.g., 0.05, the implicitly accepted tolerance level in the field) is usually considered a safeguard for reproducibility. Imposing a threshold on cluster size (perhaps combined with signal strength) to protect against the overall FPR has the undesirable trade-off cost of inflating false negative rates or type II errors, which can greatly affect individual result interpretations as well as reproducibility across studies. In general, several multiplicity-related challenges in neuroimaging appear to be tied closely to the fundamental mechanisms of NHST approaches introduced to counterbalance between type I and type II errors, which are the cornerstones of NHST. Therefore, we will take a close look at the underlying assumptions and potential problems with NHST.

Pitfalls of NHST

Following the conventional statistical procedure, the assessment for a BOLD effect is put forward through a null hypothesis H_0 as the devil's advocate; for example, an H_0 can be formulated as having no activation at a brain region under the easy condition, or as having no activation difference between the easy and difficult conditions. It is under such a null setting that statistics such as Student's t- or F-statistic are constructed, so that a standard distribution can be utilized to compute a conditional probability that is the chance of obtaining a result equal to, or more extreme than, the current outcome if H_0 is the ground truth. The rationale is that if this conditional probability is small enough, one may feel comfortable in rejecting the straw man H_0 and in accepting the alternative at a tolerable risk level.

While the above may be a reasonable formulation under some scenarios, there is a long history of arguments that emphasize the mechanical and interpretational problems with NHST (e.g., Cohen, 2014; Gelman, 2016) that might have perpetuated the reproducibility crisis across many disciplines (Loken and Gelman, 2017). For example,

i. It is a common mistake by investigators and even statistical analysts to misinterpret the conditional probability under NHST as the posterior probability of the truth of the null hypothesis (or the probability of the null event conditional on the current data at hand) even though fundamentally $P(data \mid H_0) \neq P(H_0 \mid data)$.

ii. One may conflate statistical significance with practical significance, and subsequently treat the failure to reach statistical significance as the nonexistence of any meaningful effect. It is common to read discussions in scientific literature wherein the authors implicitly (or even explicitly) treat statistically non-significant effects as if they were zero.

iii. Statistic- or p-values cannot easily be compared: the difference between a statistically significant effect and another effect that fails to pass the significance level does not necessarily itself reach statistical significance.

iv. How should the investigator handle the demarcation, due to sharp thresholding, between one effect with $p = 0.05$ (or a surviving cluster cutoff of 54 voxels) and another with $p = 0.051$ (or a cluster size of 53 voxels)?

v. The focus on statistic- or p-value seems to, in practice, lead to the preponderance of reporting only statistical, instead of effect, maps in neuroimaging, losing an effective safeguard that could have filtered out potentially spurious results (Chen et al., 2017b).

vi. One may mistakenly gain more confidence in a statistically significant result (e.g., high statistic value) in the context of data with relatively heavy noise or with a small sample size (e.g., leading to statement such as "despite the small sample size" or "despite the limited statistical power"). In fact, using statistical significance as a screener can lead researchers to make a wrong assessment about the sign of an effect or drastically overestimate the magnitude of an effect.

vii. While the conceptual classifications of false positives and false negatives make sense in a system of discrete nature (e.g., juror decision on H_0: the suspect is innocent), what are the consequences when we adopt a mechanical dichotomous approach to assessing a quantity of continuous, instead of discrete, nature?

viii. It is usually under-appreciated that the p-value, as a function of data, is a random variable, and thus itself has a sampling distribution. In other words, p-values from experiments with identical designs can differ substantially, and statistically significant results may not necessarily be replicated (Lazzeroni et al., 2016).

Within neuroimaging specifically, there are strong indications that a large portion of task-related BOLD activations are usually unidentified at the individual subject level due to the lack of power (Gonzalez-Castillo et al., 2012). The detection failure, or false negative rate, at the population level would probably be at least as large. Therefore, it is likely far-fetched to claim that no activation or no activation difference exists anywhere in the whole brain, except for the regions of white matter and cerebrospinal fluid. In other words, the global null hypothesis in enuroimaging studies is virtually never true. The situation with resting-state data analysis is likely worse than with task-related data, as the same level of noise is more impactful on seed-based correlation analysis due to the lack of objective reference effect. Since no ground truth is readily available, dichotomous inferences under NHST as to whether an effect exists in a brain region are intrinsically problematic, and it is practically implausible to truly believe the validity of H_0 as a building block when constructing a model.

Structure of the work

In light of the aforementioned backdrop, we believe that the current modeling approach is inefficient. First, we question the appropriateness of the severe penalty currently levied to the voxel- or node-wise data analysis. In addition, we endorse the ongoing statistical debate surrounding the ritualization of NHST and its dichotomous approach to results reporting and in the review process, and aim to raise the awareness of the issues embedded within NHST (Loken and Gelman, 2017) in the neuroimaging community. In addition, with the intention of addressing some of the issues discussed above, we view multiple testing as a problem of inefficient modeling induced by the conventional massively univariate methodology. Specifically, the univariate approach starts, in the same vein as a null hypothesis setting, with a pretense of spatial independence, and proceeds with many

Table 1: Acronyms and terminology.

ANOVA	analysis of variance	MCMC	Monte Carlo Markov chain
BHM	Bayesian hierarchical model	NHST	null hypothesis significance testing
FPR	false positive rate	NUTS	No-U-Turn sampler
FWE	family-wise error	power	chance of rejecting H_0 when H_0 is false
GLM	general linear model	PPC	posterior predictive check
HMC	Hamiltonian Monte Carlo	ROI	region of interest
HPD	highest posterior density	type I	chance of rejecting H_0 when H_0 is true ("false positive")
ICC	intraclass correlation	type II	failing to reject H_0 when H_0 is false ("false negative")
LME	linear mixed-effects	type M	exaggerating the effect magnitude
LOO	leave one out	type S	estimating the effect with an incorrect sign

isolated or segmented models. To avoid the severe penalty of Bonferroni correction while recovering from or compensating for the false presumption of spatial independence, the current practices deal with multiple testing by discounting the number of models due to spatial relatedness. However, the collateral damages incurred by this to-and-fro process are unavoidably the loss of modeling efficiency and the penalty for detection power under NHST.

Here, we propose a more efficient approach through BHM that could be used to confirm, complement or replace the standard NHST method. As a first step, we adopt a group analysis strategy under the Bayesian framework through hierarchical modeling on an ensemble of ROIs and use this to resolve two of the four multiplicity issues above: multiple testing and double sidedness. Those ROIs are determined independently from the current data at hand, and they can be selected through various methods such as previous studies, an anatomical or functional atlas, or parcellation of an independent dataset in a given study; the regions could be defined through masking, manual drawing, or balls about a center reported previously. The proposed BHM approach dissolves multiple testing through a hierarchical model that more accurately accounts for data structure as well as shared information, and it consequentially improves inference efficiency. The modeling approach will be extended to other scenarios in our future work.

We present this work in a purposefully (possibly overly) didactic style, reflecting our own conceptual progression. Our goal is to convert the traditional voxel-wise GLM into an ROI-based BHM through a step-wise progression of models (GLM → LME → BHM). The paper is structured as follows. In the next section, we first formulate the population analysis at each ROI through univariate GLM (parallel to the typical voxel-wise population analysis), then turn multiple GLMs into one LME by pivoting the ROIs as the levels of a random-effects factor, and lastly convert the LME to a full BHM. The BHM framework does not make statistical inference for each measuring entity (ROI in our context) in isolation. Instead, the BHM weights and borrows the information based on the precision information across the full set of entities, striking a balance between data and prior knowledge. As a practical exemplar, we apply the modeling approach to an experimental dataset and compare its performance with the conventional univariate GLM. In the Discussion section, we elaborate the advantages, limitations, and prospects of BHM in neuroimaging. Major acronyms and terms are listed in Table 1.

Theory: Bayesian hierarchical modeling

Before formally building a BHM framework, we discuss some other issues associated with NHST through two types of error that are not often discussed in neuroimaging: type S and type M. These two types of error cannot be directly captured by the FPR concept and may become severe when the effect is small relative to the noise, which is usually the situation in BOLD neuroimaging data.

Throughout this article, the word *effect* refers to a quantity of interest, usually embodied in a regression (or correlation) coefficient, the contrast between two such quantities, or the linear combination of two or more such quantities. Italic letters in lower case (e.g., α) stand for scalars and random variables; lowercase, boldfaced italic letters ($\boldsymbol{a}$) for column vectors; Roman and Greek letters for fixed and random effects in the conventional statistics context, respectively, on the righthand side of a model equation (the Greek letter θ is reserved for the effect of interest). $p(\cdot)$ represents a probability density function.

Type S and type M errors

In the NHST formulation, we formulate a null hypothesis H_0 (e.g., the effect of an easy task E is identical to a difficult one D), and then commit a type I (or false positive) error if wrongly rejecting H_0 (e.g., the effect of easy is judged to be statistically significantly different from difficult when actually their effects are the same); in contrast, we make a type II (or false negative) error when accepting H_0 when H_0 is in fact false (e.g., the effect of easy is assessed to be not statistically significant from difficult even though their effects do differ). These are the dichotomous errors associated with NHST, and the counterbalance between these two types of error are the underpinnings of typical experimental design as well results reporting.

However, we could think about other ways of framing errors when making a statistical assessment (e.g., the easy case elicits a stronger BOLD response at some region than the difficult case) conditional on the current data. We are exposed to a risk that our decision is contrary to the truth (e.g., the BOLD response to the easy condition is actually lower than to the difficult condition). Such a risk is gauged as a type S (for "sign") error when we incorrectly identify the sign of the effect; its values range from 0 (no chance of error) to 1 (full chance of error). Similarly, we make a type M (for "magnitude") error when estimating the effect as small in magnitude if it is actually large, or when claiming that the effect is large in magnitude if it is in fact small (e.g., saying that the easy condition produces a *much* large response than the difficult one when actually the difference is tiny); its values range across the positive real numbers: $[0, 1)$ correspond to underestimation of effect magnitude, 1 describes correct estimation, and $(1, \infty^+)$ mean overestimation. The two error types are illustrated in Fig. 1 for inferences made under NHST. In the neuroimaging realm, effect magnitude is certainly a property of interest, therefore the corresponding type S and type M errors would be of research interest.

Geometrically speaking, if the null hypothesis H_0 can be conceptualized as the point at zero, NHST aims at the real space $\boldsymbol{R}$ excluding zero with a pivot at the point of zero (e.g., $D - E = 0$); in contrast, type S error gauges the relative chance that a result is assessed on the wrong side of the distribution between the two half spaces of $\boldsymbol{R}$ (e.g., $D - E > 0$ or $D - E < 0$), and type M error gauges the relative magnitude of differences along segments of $\boldsymbol{R}^+$ (e.g., the ratio of *measured* to *actual* effect is $\gg 1$ or $\ll 1$). Thus, we characterize type I and type II errors as "point-wise" errors, driven by judging the equality, and describe type S and type M errors as "direction-wise," driven by the focus of inequality or directionality.

One direct application of type M error is that publication bias can lead to type M errors, as large effect estimates are more likely to filter through the dichotomous decisions in statistical inference and reviewing process. Using the type S and type M error concepts, it might be surprising for those who encounter these two error types for the first time to realize that, when the data are highly variable or noisy, or when the sample size is small with a relatively low power (e.g., 0.06), a statistically significant result at the 0.05 level is quite likely to have an incorrect sign – with a type S error rate of 24% or even higher (Gelman and Carlin, 2014). In addition, such a statistically significant result would have a type M error with its effect estimate much larger (e.g., 9 times higher) than the true value. Put it another way, if the real effect is small and sampling variance is large, then a dataset that reaches statistical significance must have an exaggerated effect estimate and the sign of the effect estimate is likely to be incorrect. Due to the ramifications of type M errors and publication filtering, an effect size from the literature could be exaggerated to some extent, seriously calling into question the usefulness of power analysis under NHST in determining sample size or power, which might explain the dramatic contrast between the common practice of power analysis as a requirement for grant applications and the reproducibility crisis across various fields. Fundamentally, power analysis inherits the same problem with NHST: a narrow emphasis on statistical significance is placed as a primary focus (Gelman and Carlin, 2013).

The typical effect magnitude in BOLD FMRI at 3 Tesla is usually small, less than 1% signal change in most

Table 2: Power, type S and type M errors estimated from simulations[a]

se	0.1			0.3			0.5			0.7			1.0		
ef	*pwr*	*S*	*M*	*pwr*	*S*	*M*	*pwr*	*S*	*M*	*pwr*	*S*	*M*	*pwr*	*S*	*M*
0.1	0.15	0.02	2.66	0.06	0.21	7.58	0.05	0.31	12.86	0.05	0.36	17.71	0.05	0.40	25.55
0.3	0.81	0.00	1.12	0.15	0.02	2.66	0.08	0.09	4.28	0.07	0.15	5.96	0.06	0.23	8.59
0.5	1.00	0.00	1.00	0.34	0.00	1.67	0.15	0.02	2.66	0.10	0.06	3.66	0.07	0.12	5.16
0.7	1.00	0.00	1.00	0.60	0.00	1.28	0.25	0.00	1.96	0.15	0.02	2.67	0.10	0.06	3.74
1.0	1.00	0.00	1.00	0.89	0.00	1.07	0.47	0.00	1.44	0.26	0.00	1.91	0.15	0.02	2.65

[a] The simulations were performed using a modified version of the code from Gelman and Carlin (2014). The power (*pwr*, gray column), type S (*S*, white column) and type M (*M*, cyan column) errors are estimated using 10,000 iterations with a Student's $t(20)$-distribution for an FMRI population analysis. The setup parameters of each simulation were the true effect (*ef*) and standard error (*se*), which are provided in the row and column labels, respectively. The true effect and standard error values ranged from 0.1-1.0, representing units of percent signal change. The combination (in purple) with effect of 0.3% and standard error of 1.0% is used to illustrate the various types of errors further in Fig. 1. In each simulation, the power is estimated as the sum of the two tailed areas beyond the threshold at the standard significance level of 0.05 (shown in blue in Fig. 1). Type S error is the ratio of the tailed area that has the opposite sign of the true effect relative to power, and type M error is expressed as the average value of "significant" results across all simulations relative to the true effect. In the lower triangle, effects with high power ($pwr \approx 1$) tend to have low type S error ($S \approx 0$) and low type M error ($M \approx 1$). However, as power decreases, type S error increases to a large fraction of unity, and $M \gg 1$.

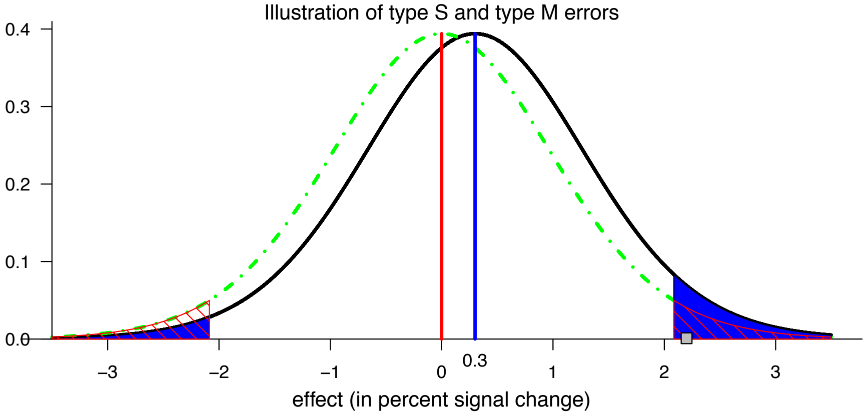


Figure 1: Illustration of the concept and interpretation for power, type I, type S and type M errors (Gelman, 2015). Suppose that there is a hypothetical Student's $t(20)$-distribution (black curve) for a true effect (blue vertical line) of 0.3 and a corresponding standard error of 1.0 percent signal change, a scenario highlighted in purple in Table 2. Under the null hypothesis (red vertical line and dot-dashed green curve), two-tailed testing with a type I error rate of 0.05 leads to having thresholds at ±2.086; FPR = 0.05 corresponds to the null distribution's total area beyond these two critical values (marked with red diagonal lines). The power is the total area of the $t(20)$-distribution for the true effect (black curve) beyond these thresholds, which is 0.06 (shaded in blue). The type S error is the ratio of the blue area in the true effect distribution's left tail beyond the threshold of -2.086 to the area in both tails, which is 23% here (i.e., the ratio of the "significant" area in the wrong-signed tail to that of the total "significant" area). If a random draw from the $t(20)$-distribution under the true effect happens to be 2.2 (small gray square), it would be identified as statistically significant at the 0.05 level, and the resulting type M error would quantify the magnification of the estimated effect size as $2.2/0.3 \approx 7.33$, which is much larger than unity.

brain regions except for areas such as motor and primary sensory cortex. Such a weak signal can be largely submerged by the overwhelming noise and distortion embedded in the FMRI data. The low power for detection of typical FMRI data analyses in typical datasets is further compounded by the modeling challenges in accurately capturing the effect. For example, even though large number of physiological confounding effects are embedded in the data, it is still difficult to properly incorporate the physiological "noises" (cardiac and respiratory effects) in the model. Moreover, habituation, saturation, or attenuation across trials or within each block are usually not considered, and such fluctuations relative to the average effect would be treated as noise or fixed- instead of random-effects (Westfall et al., 2017). There are also strong indications that a large portion of BOLD activations are usually unidentified at the individual subject level due to the lack of power (Gonzalez-Castillo et al., 2012). Because of these factors, the variance due to poor modeling overwhelms all other sources (e.g., across trials, runs, and sessions) in the total data variance (Gonzalez-Castillo et al., 2016); that is, the majority (e.g., 60-80%) of the total variance in the data is not properly accounted for in statistical models.

Achieving statistical significance has been widely used as the standard screening criterion in scientific results reporting as well as in the publication reviewing process. The difficulty in passing a commonly accepted threshold with noisy data may elicit a hidden misconception: A statistical result that survives the strict screening with a small sample size seems to gain an extra layer of strong evidence, as evidenced by phrases in the literature such as "despite the small sample size" or "despite limited statistical power." However, when the statistical power is low, the inference risks can be perilous, as demonstrated with simulations in the FMRI context in the upper triangular part of Table 2. The conventional concept of FPR controllability is not a well-balanced choice under all circumstances or combinations of effect and noise magnitudes. We consider a type S error to be more severe than a type M error, and thus we aim to control the former while at the same time reducing the latter as much as possible, parallel to the similarly lopsided strategy of strictly controlling type I errors at a tolerable level under NHST while minimizing type II errors.

Against this backdrop, we start with a classical framework, a hierarchical or multilevel model for a one-way random-effects ANOVA, and use it as a building block to expand to a Bayesian framework for neuroimaging group analysis. In evaluating this model, the controllability of inference errors will be focused on type S errors instead of the traditional FPR.

Bayesian modeling for one-way random-effects ANOVA

Suppose that there are r measured entities (e.g., ROIs), with entity j measuring the effect θ_j from n_j independent Gaussian-distributed data points y_{ij}, each of which represents a sample (e.g., trial), $i = 1, 2, ..., n_j$. The conventional statistical approach formulates r separate models,

$$y_{ij} = \theta_j + \epsilon_{ij},\ \ i = 1, 2, ..., n_j, \tag{1}$$

where ϵ_{ij} is the residual for the jth entity and is assumed to be Gaussian $\mathcal{N}(0, \sigma^2)$, $j = 1, 2, ..., r$. Depending on whether the sampling variance σ^2 is known or not, each effect can be assessed through its sample mean $\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ relative to the corresponding variance $V_j^0 = \frac{\sigma^2}{n_j}$, resulting in a Z- or t-test.

By combining the data from the r entities and further decomposing the effect θ_j into an overall effect b_0 across the r entities and the deviation ξ_j of the jth entity from the overall effect (i.e., $\theta_j = b_0 + \xi_j, j = 1, 2, ..., r$), we have a conventional one-way random-effects ANOVA,

$$y_{ij} = b_0 + \xi_j + \epsilon_{ij},\ \ i = 1, 2, ..., n_j,\ \ j = 1, 2, ..., r, \tag{2}$$

where b_0 is conceptualized as a fixed-effects parameter, ξ_j codes the random fluctuation of the jth entity from the overall mean b_0, with the assumption of $\xi_j \sim \mathcal{N}(0, \tau^2)$, and the residual ϵ_{ij} follows a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. The classical one-way random-effects ANOVA model (2) is typically formulated to examine the null hypothesis,

$$H_0 : \tau = 0, \tag{3}$$

with an F-statistic, which is constructed as the ratio of the *between* mean sums of squares and the *within* mean sums of squares. An application of this ANOVA model (2) to neuroimaging is to compute the intraclass correlation ICC(1,1) as $\frac{\tau^2}{\tau^2+\sigma^2}$ when the measuring entities are exchangeable (e.g., families with identical twins; Chen et al., 2017c).

Whenever multiple values (e.g., two effect estimates from two scanning sessions) from each measuring unit (e.g., subject or family) are correlated (e.g., the levels of a within-subject or repeated-measures factor), the data can be formulated using a linear mixed-effects (LME) model, sometimes referred to as a multilevel or hierarchical model. One natural ANOVA extension is simply to treat the model conceptually as LME, without the need of reformulating the model equation (2). However, LME can only provide point estimates for the overall effect b_0, cross-region variance τ^2 and the data variance σ^2; that is, the LME (2) cannot directly provide any information regarding the individual ξ_j or θ_j values because of over-fitting due to the fact that the number of data points is less than the number of parameters that need to be estimated.

Our interest here is neither to assess the variability τ^2 nor to calculate ICC, but instead to make statistical inferences about the individual effects θ_j. Nevertheless, the conventional NHST (3) may shed some light on potential strategies (Gelman et al., 2014) for θ_j. If the deviations ξ_j are relatively small compared to the overall mean b_0, then the corresponding F-statistic value will be small as well, leading to the decision of not rejecting the null hypothesis (3) at a reasonable, predetermined significance level (e.g., 0.05); in that case, we can estimate the equal individual effects θ_j using the overall weighted mean $\bar{y}_{\cdot\cdot}$ through full pooling with all the data,

$$\hat{\theta}_1 = \hat{\theta}_2 = ... = \hat{\theta}_r = \bar{y}_{\cdot\cdot} = \frac{\sum_{j=1}^{r} \frac{1}{\sigma_j^2} \bar{y}_{\cdot j}}{\sum_{j=1}^{r} \frac{1}{\sigma_j^2}}, \tag{4}$$

where $\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ and $\sigma_j^2 = \frac{\sigma^2}{n_j}$ are the sampling mean and variance for the jth measuring entity, and the subscript dot $(\cdot)$ notation indicates the (weighted) mean across the corresponding index(es). On the other hand, if the deviations ξ_j are relatively large, so is the associated F-statistic value, leading to the decision of rejecting the null hypothesis (3); similarly, we can reasonably estimate θ_j with no pooling across the r entities; that is, each θ_j is estimated using the jth measuring entity's data separately,

$$\hat{\theta}_j = \bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \ j = 1, 2, ..., r. \tag{5}$$

However, in estimating θ_j we do not have to take a dichotomous approach of choosing, based on a preset significance level, between these two extreme choices, the overall weighted mean $\bar{y}_{\cdot\cdot}$ in (4) through full pooling and the separate means $\bar{y}_{\cdot j}$ in (4) with no pooling; instead, we could make the assumption that a reasonable estimate to θ_j lies somewhere along the continuum between $\bar{y}_{\cdot\cdot}$ and $\bar{y}_{\cdot j}$, with its exact location derived from the data instead of by imposing an arbitrary threshold. This thinking brings us to the Bayesian methodology.

To simplify the situation, we first assume a known sampling variance σ^2 for the ith data point (e.g., trial) for the jth entity; or, in Bayesian-style formulation, we build a BHM about the distribution of y_{ij} conditional on θ_j,

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2), \ i = 1, 2, ..., n_j, \ j = 1, 2, ..., r. \tag{6}$$

With a prior distribution $\mathcal{N}(b_0, \tau^2)$ for θ_j and a noninformative uniform hyperprior for b_0 given τ (i.e., $b_0|\tau \sim 1$), the conditional posterior distributions for θ_j can be derived (Gelman et al., 2014) as,

$$\theta_j|b_0, \tau, y \sim \mathcal{N}(\hat{\theta}_j, V_j), \text{ where } \hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}b_0}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}},\ V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}},\ \sigma_j^2 = \frac{\sigma^2}{n_j},\ j = 1, 2, .., r. \tag{7}$$

The analytical solution (7) indicates that $\frac{1}{V_j} = \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}$, manifesting an intuitive fact that the posterior precision is the cumulative effect of the data precision and the prior precision; that is, the posterior precision is improved by the amount $\frac{1}{\tau^2}$ relative to the data precision $\frac{1}{\sigma_j^2}$. Moreover, the expression for the posterior mode of $\hat{\theta}_j$ in (7) shows that the estimating choice in the continuum can be expressed as a precision-weighted average between the individual sample means $\bar{y}_{\cdot j}$ and the overall mean b_0:

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}b_0}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} = w_j\bar{y}_{\cdot j} + (1 - w_j)b_0 = b_0 + w_j(\bar{y}_{\cdot j} - b_0) = \bar{y}_{\cdot j} - (1 - w_j)(\bar{y}_{\cdot j} - b_0),\ j = 1, 2, .., r, \tag{8}$$

where the weights $w_j = \frac{V_j}{\sigma_j^2}$. The precision weighting in (8) makes intuitive sense in terms of the previously described limiting cases:

i. The full pooling (4) corresponds to $w_j = 0$ or $\tau^2 = 0$, which means that the θ_j are assumed to be the same or fixed at a common value.

ii. The no pooling (5) corresponds to $w_j = 1$ or $\tau^2 = \infty$, indicating that the r effects θ_j are uniformly distributed within $(-\infty, \infty)$; that is, it corresponds to a noninformative uniform prior on θ_j.

iii. The partial pooling (7) or (8) reflects the fact that the r effects θ_j are *a priori* assumed to follow an independent and identically distribution, the prior $\mathcal{N}(b_0, \tau^2)$. Under the Bayesian framework, we make statistical inferences about the r effects θ_j with a posterior distribution (7) that includes the conventional dichotomous decisions between full pooling (4) and no pooling (5) as two special and extreme cases. Moreover, as expressed in (8), the Bayesian estimate $\hat{\theta}_j$ can be conceptualized as the precision-weighted average between the individual estimate $\bar{y}_{\cdot j}$ and the overall (or prior) mean b_0, the adjustment of θ_j from the overall mean b_0 toward the observed mean $\bar{y}_{\cdot j}$, or conversely, the observed mean $\bar{y}_{\cdot j}$ being shrunk toward the overall mean b_0.

An important concept for a Bayesian model is exchangeability. Specifically for the BHM (6), the effects θ_j are exchangeable if their joint distribution $p(\theta_1, \theta_2, ..., \theta_r)$ is immutable or invariant to any random permutation among their indices or orders (e.g., $p(\theta_1, \theta_2, ..., \theta_r)$ is a symmetric function). Using the ROIs as an example, exchangeability means that, without any *a priori* knowledge about their effects, we can randomly shuffle or relabel them without reducing our knowledge about their effects. In other words, complete ignorance equals exchangeability: before poring over the data, there is no way for us to distinguish the regions from each other. When the exchangeability assumption can be assumed, the corollary is that the effects θ_j are *a priori* independently and identically distributed (*iid*), as shown in the derivation of the posterior distribution (8) from the prior distribution $\mathcal{N}(b_0, \tau^2)$ for θ_j.

To complete the Bayesian inferences for the model (6), we proceed to obtain (i) $p(b_0, \tau|y)$, the marginal posterior distribution of the hyperparameters (b_0, τ), (ii) $p(b_0|\tau, y)$, the posterior distribution of b_0 given τ, and (iii) $p(\tau|y)$, the posterior distribution of τ with a prior for τ, for example, a noninformative uniform distribution $p(\tau) \sim 1$. In practice, the numerical solutions are achieved in a backward order, through Monte Carlo simulations of τ to get $p(\tau|y)$, simulations of b_0 to get $p(b_0|\tau, y)$, and, lastly, simulations of θ_j to get $p(\theta_j|b_0, \tau, y)$ in (7).

Assessing type S error under BHM

In addition to the advantage of information merging across the r entities between the limits of complete and no pooling, a natural question remains: how does BHM perform in terms of the conventional type I error as well as type S and type M errors? With the "standard" analysis of r separate models in (1), each effect θ_j is assessed against the sampling variance $V_j^0 = \sigma_j^2$. In contrast, under the BHM (6), the posterior variance, as shown in (7), is $V_j = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}$, $\sigma_j^2 = \frac{\sigma^2}{n_j}$. As the ratio of the two variances $\frac{V_j^0}{V_j} = \frac{\tau^2}{\tau^2+\sigma_j^2}$ is always less than 1 (except for the limiting cases of $\sigma^2 \to 0$ or $\tau^2 \to \infty$), BHM generally assigns a larger uncertainty than the conventional approach with no pooling. That is, the inference for each effect θ_j based on the unified model (6) is more conservative than when the effect is assessed individually through the model (1). Instead of tightening the overall FPR through some kind of correction for multiplicity among the r separate models, BHM addresses the multiplicity issue through precision adjustment or partial pooling under one model with a shrinking or pooling strength of $\sqrt{\frac{V_j^0}{V_j}} = \frac{1}{\sqrt{1+\sigma_j^2/\tau^2}}$.

Simulations (Gelman and Tuerlinckx, 2014) indicate that, when making inference based on the 95% quantile interval of the posterior distribution for a single effect θ_j (j is fixed, e.g., $j = 1$), the type S error rate for the Bayesian model (6) is less than 0.025 under all circumstances. In contrast, the conventional model (1) would have a substantial type S error rate especially when the sampling variance is large relative to the cross-entities variance (e.g., $\sigma_j^2/\tau^2 > 2$); specifically, type S error reaches 10% when $\sigma_j^2/\tau^2 = 2$, and may go up to 50% if σ_j^2 much larger than τ^2. When multiple comparisons are performed, a similar patterns remains; that is, the type S error rate for the Bayesian model is in general below 2.5%, and is lower than the conventional model with rigorous correction (e.g., Tukey's honestly significant difference test, wholly significant differences) for multiplicity when $\sigma/\tau > 1$. The controllability of BHM on type S errors is parallel to the usual focus on type I errors under NHST; however, unlike NHST in which the typical I error rate is deliberately controlled through a an FPR threshold, the controllability of type S errors under BHM is intrinsically embedded in the modeling mechanism without any explicit imposition.

The model (6) is typically seen in Bayesian statistics textbooks as an intuitive introduction to BHM (e.g., Gelman et al., 2014). With the indices i and j coding the task trials and ROIs, respectively, the ANOVA model (2) or its Bayesian counterpart (6) can be utilized to make inferences on an ensemble of ROIs at the individual subject level. The conventional analysis would have to deal with the multiplicity issue because of separate inferences at each ROI (i.e., entity). In contrast, there is only one integrated model (6) that leverages the information among the r entities, and the resulting partial pooling effectively dissolves the multiple testing concern. However, the modeling framework can only be applied for single subject analysis, and it is not suitable at the population level; nevertheless, it serves as an intuitive tool for us to move forward to more sophisticated scenarios. As our main focus here is population analysis, we extend the approach further to a two-way ANOVA structure, and elucidate the advantages of data calibration and partial pooling in more details.

Bayesian modeling for two-way random-effects ANOVA

At the population level, the variability across n subjects has to be accounted for; in addition, the within-subject correlation structure among the ROIs also needs to be maintained. The conventional approach formulates r separate GLMs each of which fits the data from the ith subject at the jth ROI,

$$y_{ij} = \theta_j + \epsilon_{ij},\ i = 1, 2, ..., n, \tag{9}$$

where $j = 1, 2, ..., r$, ϵ_{ij} is the residual term that is assumed to independently and identically follow $\mathcal{N}(0, \sigma^2)$. Each of the r models in (9) essentially corresponds to a Student's t-test, and the immediate challenge is the multiple testing issue among those r models: with the assumption of exchangeability among the ROIs, is Bonferroni correction the only valid solution? If so, most neuroimaging studies would have difficulty in adopting ROI-based analysis due to this severe penalty, which may be the major reason that discourages the use of region-level analysis with a large number of regions.

We first extend the one-way random-effects ANOVA model (2) to a two-way random-effects ANOVA, and formulate the following platform with data from n subjects,

$$y_{ij} = b_0 + \pi_i + \xi_j + \epsilon_{ij},\ i = 1, 2, ..., n,\ j = 1, 2, ..., r, \tag{10}$$

where π_i and ξ_j code the deviation or random effect of the ith subject and jth ROI from the overall mean b_0, respectively, and they are assumed to be *iid* with $\mathcal{N}(0, \lambda^2)$ and $\mathcal{N}(0, \tau^2)$, and ϵ_{ij} is the residual term that is assumed to follow $\mathcal{N}(0, \sigma^2)$.

Parallel to the situation with the one-way ANOVA (2), the two-way ANOVA (10) can be conceptualized as an LME without changing its formulation. Specifically, the overall mean b_0 is a fixed-effects parameter, while both the subject- and ROI-specific effects, π_i and ξ_j, are treated as random variables. In addition, we continue to define $\theta_j = b_0 + \xi_j$ as the effect of interest at the jth ROI. The LME framework has been well developed over the past half century, under which we can estimate variance components such as λ^2 and τ^2, and fixed effects such as b_0 in (10). Therefore, conventional inferences can be made by constructing an appropriate statistic for a null hypothesis. Its modeling applicability and flexibility have been substantiated by its adoption in FMRI group analysis (Chen et al., 2013). Furthermore, the LME formulation (10) has a special structure, a crossed random-effects structure, which has been applied to inter-subject correlation (ISC) analysis for naturalistic scanning (Chen et al., 2017a) and to ICC analysis for ICC(2,1) (Chen et al., 2017c).

However, LME cannot offer a solution in making inferences regarding the ROI effects θ_j: to estimate θ_j, the LME (10) would become over-parameterized (i.e., an over-fitting problem). To proceed for the sake of intuitive interpretations, we temporarily assume a known sampling variance σ^2, a known cross-subjects variance λ^2, and a known cross-ROI variance τ^2, and transform the ANOVA (10) to its Bayesian counterpart,

$$y_{ij} | \pi_i, \theta_j \sim \mathcal{N}(\pi_i + \theta_j, \sigma^2),\ i = 1, 2, ..., n,\ j = 1, 2, ..., r. \tag{11}$$

Then the posterior distribution of θ_j with prior distributions, $\pi_i \sim \mathcal{N}(0, \lambda^2)$ and $\theta_j \sim \mathcal{N}(b_0, \tau^2)$, can be analytically derived with the data $\boldsymbol{y} = \{y_{ij}\}$ (Appendix A),

$$\theta_j | b_0, \tau, \lambda, \boldsymbol{y} \sim \mathcal{N}(\hat{\theta}_j, V),\ \text{where } \hat{\theta}_j = \frac{\frac{n}{\lambda^2+\sigma^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}b_0}{\frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}},\ V = \frac{1}{\frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}},\ j = 1, 2, .., r. \tag{12}$$

Similarly to the previous section, we have an intuitive interpretation for $\frac{1}{V} = \frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}$: the posterior precision for $\theta_j | b_0, \tau, \lambda, \boldsymbol{y}$ is the sum of the cross-ROI precision $\frac{1}{\tau^2}$ and the combined sampling precision $\frac{n}{\lambda^2+\sigma^2}$. Under the r completely separate GLMs in (9), the cross-subjects variance λ^2 and the sampling variance σ^2 could not be estimated separately. Interestingly, the following relationship,

$$\frac{n}{\lambda^2 + \sigma^2} < \frac{1}{V} = \frac{n}{\lambda^2 + \sigma^2} + \frac{1}{\tau^2} \leq \frac{n}{\sigma^2} + \frac{1}{\tau^2}, \tag{13}$$

reveals that the posterior precision lies somewhere among the precisions of $\hat{\theta}_j$ from the r separate GLMs. Furthermore, the posterior mode of $\hat{\theta}_j$ in (12) can be expressed as a weighted average between the individual sample means $\bar{y}_{\cdot j}$ and the overall mean b_0,

$$\hat{\theta}_j = \frac{\frac{n}{\lambda^2+\sigma^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}b_0}{\frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}} = w\bar{y}_{\cdot j} + (1-w)b_0 = b_0 + w(\bar{y}_{\cdot j} - b_0) = \bar{y}_{\cdot j} - (1-w)(\bar{y}_{\cdot j} - b_0),\ j = 1, 2, .., r, \tag{14}$$

where the weight $w = \frac{nV}{\lambda^2+\sigma^2}$, indicating the counterbalance of partial pooling between the individual mean $\bar{y}_{\cdot j}$ for the jth entity and the overall mean b_0, the adjustment of θ_j from the overall mean b_0 toward the observed mean $\bar{y}_{\cdot j}$, or the observed mean $\bar{y}_{\cdot j}$ being shrunk toward the overall mean b_0.

Related to the concept of ICC, the correlation between two ROIs, j_1 and j_2, due to the fact that they are measured from the same set of subjects, can be derived in a Bayesian fashion as,

$$\begin{aligned} corr(y_{ij_1}, y_{ij_2}|\lambda^2, \tau^2, \sigma^2) &= \frac{cov(\pi_i + \theta_{j_1} + \epsilon_{ij_1}, \pi_i + \theta_{j_2} + \epsilon_{ij_2})}{\sqrt{var(\pi_i + \theta_{j_1} + \epsilon_{ij_1}) var(\pi_i + \theta_{j_2} + \epsilon_{ij_2})}}|\lambda^2, \tau^2, \sigma^2 \\ &= \frac{\lambda^2}{\lambda^2 + \tau^2 + \sigma^2},\ j_1, j_2 = 1, 2, .., r\ (j_1 \neq j_2). \end{aligned} \tag{15}$$

Similarly, the correlation between two subjects, i_1 and i_2, due to the fact that their effects are measured from the same set of ROIs, can be derived in a Bayesian fashion as,

$$\begin{aligned} corr(y_{i_1 j}, y_{i_2 j}|\lambda^2, \tau^2, \sigma^2) &= \frac{cov(\pi_{i_1} + \theta_j + \epsilon_{i_1 j}, \pi_{i_2} + \theta_j + \epsilon_{i_2 j})}{\sqrt{var(\pi_{i_1} + \theta_j + \epsilon_{i_1 j}) var(\pi_{i_2} + \theta_j + \epsilon_{i_2 j})}}|\lambda^2, \tau^2, \sigma^2 \\ &= \frac{\tau^2}{\lambda^2 + \tau^2 + \sigma^2},\ i_1, i_2 = 1, 2, .., n\ (i_1 \neq i_2). \end{aligned}$$

The exchangeability assumption is crucial here as well for the BHM system (10). Conditional on ξ_j (i.e., when the ROI is fixed at index j), the subject effects π_i can be reasonably assumed to be exchangeable since the experiment participants are usually recruited randomly from a hypothetical population pool as representatives (thus the concept of coding them as dummy variables). As for the ROI effects ξ_j, here we simply assume the validity of exchangeability conditional on the subject effect π_i (i.e., when subject is fixed at index i), and address the validity later in Discussion.

So far, we have presented a "simplest" BHM scenario. Specifically, we have: ignored the possibility of incorporating any explanatory variables such as subject-specific quantities (e.g., age, IQ) or behavioral data (e.g., reaction time); assumed known variances such as τ^2 and σ^2; and presumed that the data y_{ij} have been directly measured without precision information available. Further extensions are needed and discussed for realistic applications in the next subsection.

Further extensions of Bayesian modeling for two-way random-effects ANOVA and full Bayesian implementations

To gain intuitive interpretations, we have so far assumed that the variance σ^2 in (6) and the variances σ^2, λ^2 and τ^2 in (11) are known. In practice, those parameters for the prior distributions are not available. Approximate (or empirical) Bayesian approaches could be adopted to provide a computationally economical "workaround" solution. For example, one possibility is to first solve the corresponding LME and directly apply the estimated variances to the analytical formulas (7) or (12). However, there are two limitations that are associated with approximate Bayesian approaches. The reliability or uncertainty for the estimated variances are not taken into consideration and thus may result in inaccurate posterior distributions. In addition, analytical formulas such as (7) and (12) are usually not available when we extend the prototypical models (6) and (11) to more generic scenarios, as shown below.

At the population level, one may incorporate one or more subject-specific covariates such as subject-grouping variables (patients vs. controls, genotypes, adolescents vs. adults) and/or quantitative explanatory variables (age, behavioral or biometric data). To be able to adapt such scenarios, we first need to expand the models considered previously with a simple intercept (Student's t-test) to r separate GLMs, generalizing the model (9),

$$y_{ij} = \boldsymbol{x}^T \boldsymbol{\theta}_j + \epsilon_{ij},\ i = 1, 2, ..., n, \tag{16}$$

where the vector $\boldsymbol{x}$ contains the subject-specific values of the covariates, with the first component 1 that is associated with the intercept, and the vector $\boldsymbol{\theta}_j$ codes the effects associated with the covariates in $\boldsymbol{x}$, $j = 1, 2, ..., r$.

In parallel, the conventional two-way random-effects ANOVA or LME (10) evolves to

$$y_{ij} = \boldsymbol{x}^T \boldsymbol{b} + \pi_i + \boldsymbol{x}^T \boldsymbol{\xi}_j + \epsilon_{ij},\ i = 1, 2, ..., n,\ j = 1, 2, ..., r, \tag{17}$$

where $\boldsymbol{b}$ and $\boldsymbol{\xi}_j$ represent the population effects and subject deviations corresponding to those covariates, respectively. Similarly, the BHM counterpart can be formulated as,

$$y_{ij}|\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{\xi}_j \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{b} + \boldsymbol{x}^T \boldsymbol{\xi}_j, \lambda^2 + \sigma^2),\ i = 1, 2, ..., n,\ j = 1, 2, ..., r. \tag{18}$$

Under the BHM (18), the effect of interest θ_j can be an element of $\boldsymbol{b}$, the intercept (as in (11)) or the effect for one of the covariates in $\boldsymbol{x}$. When there is only one covariate x, the three models (16), (17) and (18) simplify to, respectively,

$$y_{ij} = \theta_{0j} + \theta_{1j} x + \epsilon_{ij},\ i = 1, 2, ..., n, \tag{19}$$

$$y_{ij} = b_0 + b_1 x + \pi_i + \xi_{0j} + \xi_{1j} x + \epsilon_{ij},\ i = 1, 2, ..., n,\ j = 1, 2, ..., r, \tag{20}$$

$$y_{ij}|x, b_0, b_1, \xi_{0j}, \xi_{1j} \sim \mathcal{N}(b_0 + b_1 x + \xi_{0j} + \xi_{1j} x, \lambda^2 + \sigma^2),\ i = 1, 2, ..., n,\ j = 1, 2, ..., r. \tag{21}$$

The discussion so far has assumed that data y_{ij} are directly collected without measurement errors. However, in some circumstances (including neuroimaging) the data are summarized through one or more analytical steps. For example, the data y_{ij} in FMRI can be the BOLD responses from subjects under a condition or task that are estimated through a time series regression model, and the estimates are not necessarily equally reliable. Therefore, a third extension is desirable to broaden our model (18) so that we can accommodate the situation where the separate variances σ_{ij}^2 of measurement errors for each ROI and subject are known and should be included in the model (18) as inputs, instead of being treated as one hyperparameter. Similarly to the conventional meta-analysis, a BHM with known sampling variances can be effectively analyzed by simply treating the variances as known values.

Numerical implementations of BHM

Since no analytical formula is generally available for the BHM (18), we proceed with the full Bayesian approach hereafter, and adopt the algorithms implemented in Stan, a probabilistic programming language and a math library in C++ on which the language depends (Stan Development Team, 2017). In Stan, the main engine for Bayesian inferences is No-U-Turn sampler (NUTS), a variant of Hamiltonian Monte Carlo (HMC) under the category of gradient-based Markov chain Monte Carlo (MCMC) algorithms.

Some conceptual and terminological clarifications are warranted here. Under the LME framework, the differentiation between fixed- and random-effects is clearcut: fixed-effects parameters (e.g., $\boldsymbol{b}$ in (17)) are considered universal constants at the population level to be estimated; in contrast, random-effects variables (e.g., $\boldsymbol{\xi_j}$ in (17)) are assumed to be random and follow a presumed distribution. In contrast, there is no such distinction between fixed and random effects in Bayesian formulations, and all effects are treated as parameters and are assumed to have prior distributions. Nevertheless, there is a loose correspondence between LME and BHM: fixed effects under LME are usually termed as population effects under BHM, while random effects in LME are typically referred to as entity effects[1] under BHM.

Essentially, the full Bayesian approach for the BHM systems (6), (11), and (18) can be conceptualized as assigning hyperpriors to the parameters in the LME or ANOVA counterparts (2), (10), and (17). Our prior distribution choices follow the general recommendations in Stan (Stan Development Team, 2017). Regarding

[1] Entity effects are more popularly called group effects in the Bayesian literature. However, to avoid potential confusions with the neuroimaging terminology in which the word *group* refers to subject categorization (e.g., males vs. females, patients vs. controls) or the analytical step of generalization from individual subjects (corresponding to the word *population* in the Bayesian literature), we adopt *entity* to mean each measuring entity such as subject and ROI in the current context.

hyperpriors, an improper flat (noninformative uniform) distribution over the real domain for the population parameters (e.g., $\boldsymbol{b}$ in (18)) is adopted, since we usually can afford the vagueness thanks to the usually satisfactory amount of information available in the data at the population level. For the scaling parameters at the entity level, the variance-covariance matrix for $\boldsymbol{\xi_j}$ and the variance of the residuals ϵ_{ij} in (18), we use a weakly informative prior such as a Student's half-$t(3, 0, 1)$[2] or half-Gaussian $\mathcal{N}(0, 1)$ (restricting to the positive half of the respective distribution). For the covariance structure in $\boldsymbol{\xi_j}$ of BHM (18), the LKJ correlation prior[3] is used with the parameter $\zeta = 1$ (i.e., jointly uniform over all correlation matrices of the respective dimension). Lastly, the variance for the residuals ϵ_{ij} in (18) is assigned with a half Cauchy prior with a scale parameter depending on the standard deviation of y_{ij}.

Bayesian inference hinges around the whole posterior distribution. For practical considerations in results reporting, modes such as mean and median are typically used to show the centrality, while a percentage (e.g., 95%) quantile interval or highest posterior density (HPD) provides a condensed and practically useful summary of the posterior distribution. The typical workflow to obtain the posterior distribution for an effect of interest is the following. Multiple (e.g., 4) Markov chains are usually run in parallel with each of them going through a predetermined number (e.g., 2000) of iterations, half of which are thrown away as warm-up (or "burn-in") iterations while the rest are used as random draws from which posterior distributions are derived. To gauge the consistency of an ensemble of Markov chains, the split $\hat{R}$ statistic (Gelman et al., 2014) is provided as a potential scale reduction factor on split chains and as a diagnostic parameter to assist the analyst in assessing the quality of the chains. Ideally, fully converged chains correspond to $\hat{R} = 1.0$, but in practice $\hat{R} < 1.1$ is considered acceptable. Another useful parameter, the number of effective sampling draws after warm-up, measures the number of independent draws from the posterior distribution that would be expected to produce the same standard deviation of the posterior distribution as is calculated from the dependent draws from HMC. As the sampling draws are not always independent with each other, especially when Markov chains proceed slowly, one should make sure that the effective sample size is large enough relative to the total sampling draws so that a reasonable accuracy can be achieved to derive the quantile intervals for the posterior distribution. For example, a 95% quantile interval requires at least an effective sample size of 200. As computing parallelization can only be executed for multiple chains of the HMC algorithms, the typical BHM analysis can be effectively conducted on any system with 4 or more CPUs.

One important aspect of the Bayesian framework is model quality check through various prediction accuracy metrics. The aim of the quality check is not to reject the model, but rather to check whether it fits the data well. For instance, posterior predictive check (PPC) simulates replicated data under the fitted model and then graphically compares actual data y_{ij} to the model prediction. The underlying rationale is that, through drawing from the posterior predictive distribution, a reasonable model should generate new data that look similar to the acquired data at hand. As a model validation tool, PPC intuitively provides a visual tool to examine any systematic differences and potential misfit of the model, similar to the visual examination of plotting a fitted regression model against the original data. Leave-one-out (LOO) cross-validation using Pareto-smoothed importance sampling (PSIS) is another accuracy tool (Vehtari et al., 2017) that uses probability integral transformation (PIT) checks through a quantile-quantile (Q-Q) plot to compare the LOO-PITs to the standard uniform or Gaussian distribution.

BHM applied to an ROI-based group analysis

To demonstrate the performances of BHM in comparison to the conventional univariate approach at the ROI level, we utilize an experimental dataset from a previous FMRI study (Xiao et al., 2017). Briefly, a cohort of 124

[2] See https://en.wikipedia.org/wiki/Folded-t_and_half-t_distributions for the density $p(\nu, \mu, \sigma^2)$ of folded non-standardized t-distribution, where the parameters ν, μ, and σ^2 are the degrees of freedom, mean, and variance.

[3] The LKJ prior (Lewandowski, Kurowicka, and Joe, 2009) is a distribution over symmetric positive-definite matrices with the diagonals of 1s.

Table 3: ROIs and FWE correction for their associated clusters[a]

voxel-wise p	cluster threshold	number of surviving ROIs	ROIs
0.001	28	2	R PCC, PCC/PrC
0.005	66	4	R PCC, PCC/PrC., L IPL, L TPJ
0.01	106	4	R PCC, PCC/PrC., L IPL, L TPJ
0.05	467	4	R PCC, PCC/PrC., L IPL, L TPJ
0.05*	467	(4)	(L aMTS/aMTG, R TPJp, vmPFC, dmPFC)

[a] Monte Carlo simulations were conducted using a mixed exponential spatial autocorrelation function (Cox et al., 2017) instead of FWHM to determine the cluster threshold (voxel size: $3 \times 3 \times 3$ mm^3). The ROI abbreviations are listed in Table 4.
*Special note for the last row (voxel-wise p-value of 0.05): four ROIs including L IPL, L TPJ, R PCC, PCC/PrC survived together with their clusters from the FWE correction, and the other four ROIs listed here (L aMTS/aMTG, R TPJp, vmPFC, and dmPFC) did not survive with their clusters but showed some evidence of effect when the cluster size requirement was dropped.

typically developing children (mean age = 6.61 years, SD = 1.41 years, range = 4 to 8.93 years; 54 males) was scanned using naturalistic FMRI. In addition, a subject-level covariate was included in the analysis: the overall theory of mind ability based on a parent-report measure (the theory of mind inventory, or ToMI). FMRI images were acquired while the children watched Inscapes videos with the following EPI scan parameters: $B_0 = 3$ T, flip angle = 70 °, echo time = 25 ms, repetition time = 2000 ms, 36 slices, planar field of view = 192 × 192 mm^2, voxel size = 3.0 × 3.0 × 3.5 mm^3, 210 volumes with a total scanning time of 7 minutes and 6 seconds. Twenty-one ROIs (Table 4) were selected for their potential relevancy to the study, and mean Fisher-transformed z-scores were extracted at each ROI from the output of seed-based correlation analysis (seed: right temporo-parietal junction at the MNI coordinates of (50, -60, 18)) from each of the 124 subjects. The effect of interest at the population level is the correlation between the behavioral measure of the overall ToMI and the association/correlation with the seed. A whole brain analysis showed the difficulty of some clusters surviving FWE correction (Table 3).

The data from the 21 ROIs were analyzed through the modeling triplets, GLM (19), LME (20) and BHM (21), with the effect of interest at each ROI being the association between ToMI and the correlation with the seed: $\theta_{1j} = b_1 + \xi_{1j}$. The exchangeability assumption for LME and BHM was deemed reasonable because, prior to the analysis, no specific information was available regarding the order and relatedness of the effects across subjects and ROIs. It is worth noting that the data were skewed with a longer right tail than left (black solid curve in Fig. 3a and Fig. 3b). When fitted at each ROI separately with GLM (simple regression in this case) using the overall ToMI as an explanatory variable, the model yielded lackluster fitting (Fig. 3a) in terms of skewness, the two tails, and the peak area. As shown in Table 6, five ROIs (R PCC, R TPJp, L IPL, L TPJ, and L aMTS/aMTG) reached a two-tailed significance level of 0.05, and two ROIs (PCC/PrC and vmPFC) achieved a two-tailed significance level of 0.1 (or one-tailed significance level of 0.05 if directionality was *a priori* known). As indicated in Fig. 2, a substantial amount of heterogeneity existed on the sampling variances among ROIs. However, the burden of FWE correction (e.g., Bonferroni) for the ROI-based approach with univariate GLM is so severe that none of the ROIs could survive the penalizing metric.

The ROI data were fitted with LME (20) and BHM (21) using the overall ToMI as an explanatory variable using, respectively, the R (R Core Team, 2017) package lme4 (Bates et al., 2015) and Stan with the code translated to C++ and compiled. Runtime for BHM was 5 minutes including approximately 1 minute of code compilation on a Linux system (Fedora 25) with AMD Opteron 6376 at 1.4 GHz. All the parameter estimates were quite similar between the two models (Table 5), although priors were incorporated into BHM. Compared to the traditional ROI-based GLM, the shrinkage under BHM can be seen in Table 6 and Fig. 2: most effect estimates were dragged toward the center; the relationship for the posterior variance, (13), is illustrated in Fig. 2. Similar to the ROI-based GLM without correction, BHM demonstrated (Table 6) strong evidence within 95% HPD interval of the overall ToMI effect at six ROIs (R PCC, R TPJp, L IPL, PCC/PrC, L TPJ, and L aMTS/aMTG), and 90% HPD interval (or 95% HPD interval if directionality was *a priori* known) at two ROIs (dmMPFC and vmPFC).

Table 4: MNI coordinates of the 21 ROIs[a]

No	ROI	Coordinates (x, y, z)
1	R PCC	(8, -59, 35)
2	R TPJp	(56,-56,25)
3	R Insula	(49, -8, -11)
4	L IPL	(-55, -65, 27)
5	L SFG	(-7, 58, 21)
6	R IFG (BA45)	(47, 22, 6)
7	R IFG (BA9)	(60, 25, 19)
8	L MTG	(-51, -62, 5)
9	L CG	(-5, 8, 42)
10	L IFG	(-46, 24, 7)
11	ACC	(0, 38, 10)
12	SGC	(-2, 32, -8)
13	PCC/PrC	(-2, -52, 26)
14	dmPFC	(-2, 5, 14)
15	L TPJ	(-46 -66, 18)
16	L vBG	(-6 ,10, -8)
17	R vBG	(6, 10, -8)
18	L aMTS/aMTG	(-54, -10, -20)
19	R Amy/Hippo	(24, -8, -22)
20	L Amy/Hippo	(-24, -10, -20)
21	vmPFC	(-2, 50, -10)

[a] Each ROI was created as a ball with a center at the coordinates (in millimeters) from the literature (Xiao et al., 2017) and a radius of 6 mm. ROI abbreviations: L, left hemisphere; R, right hemisphere; PCC/PrC, precuneus/posterior cingulate cortex; TPJp, posterior temporo-parietal junction; IPL, inferior parietal lobe; SFG, superior frontal gyrus; IFG, inferior frontal gyrus; aMTS/aMTG, anterior middle temporal sulcus/gyrus; CG, cingulate gyrus; ACC, anterior cingulate cortex; SGC, subgenual cingulate cortex; dmPFC, dorsomedial prefrontal cortex; vBG, ventral basal ganglia; Amy/Hippo, amygdala/hippocampus; vmPFC, ventromedial prefrontal cortex.

One exception to the general shrinkage under BHM is that the mean effect, 0.025, at the region of R TPJp (second row in Table 6) was actually higher than that under GLM, 0.018. Such an exception occurred because the final result is a combination or a tug of war between the shrinkage impact as shown in (14) and the correlation structure among the ROIs as shown in (15). Noticeably, the quality and fitness of BHM can be diagnosed and verified through posterior predictor check (Fig. 3a and Fig. 3b) that compares the observed data with the simulated data based on the model: not only did the BHM accommodate the skewness of the data better than GLM, but also did the partial pooling render much better fit for the peak and both tails as well. Cross validation through LOO (Fig. 3c and Fig. 3d) also manifested the advantage of BHM fitting over GLM. Nevertheless, there is still room for the improvement of BHM: the peak area could be fitted better, which may require nonlinearity or incorporating other potential covariates.

One apparent aspect that the ROI-based BHM excels is the completeness and transparency in results reporting: if the number of ROIs is not overwhelming (e.g., less than 100), the summarized results for every ROI can be completely presented in a tabular form (*c.f.* Table 6) despite their differential strength of statistical evidence, unlike the whole brain analysis in which the results are typically reported as the tips of icebergs above the water. In addition, one does not have to stick to a single harsh thresholding when deciding a criterion on the ROIs for discussion; for instance, even if an ROI lies outside of, but close to, the 95% HPD interval (e.g., dmMPFC and vmPFC in Table 6), it can still be reported and discussed as long as all the details are revealed. Such flexibility and transparency are difficult to navigate or maneuver through cluster thresholding at the whole brain level. As a counterpart to NHST, a p-value could be provided for each effect under BHM in the sense as illustrated in Table 7; however, we opt not to do so for two reasons: 1) such a p-value could be easily misinterpreted in the NHST sense, and, more importantly, 2) it is the predictive intervals shown in Table 6, not the single p-values, that characterize the posterior distribution, providing more information than just binary ("in or out") thresholding.

Those four regions (L IPL, L TPJ, R PCC, PCC/PrC) that passed the FWE correction at voxel-wise p-cutoff

a. LME result with model (20):

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 subject  (Intercept) 5.816e-03 0.076264
 ROI      (Intercept) 2.340e-02 0.152974
          ToMI        6.397e-05 0.007998 0.88
 Residual             2.341e-02 0.153008
Number of obs: 2604, groups:  subject, 124; ROI, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept) 0.168125   0.034209   4.915
ToMI        0.006863   0.004000   1.716

Correlation of Fixed Effects:
      (Intr)
total 0.373
```

b. BHM result with model (21):

```
Group-Level Effects:
~ROI (Number of levels: 21)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.16      0.03     0.12     0.23        551 1.00
sd(ToMI)                 0.01      0.00     0.00     0.01        947 1.00
cor(Intercept,total)     0.77      0.16     0.37     0.99       1054 1.00

~subject (Number of levels: 124)
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)     0.08      0.01     0.07     0.09        500 1.01

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept     0.17      0.04     0.09     0.24        162 1.03
ToMI          0.01      0.00    -0.00     0.01        468 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     0.15      0.00     0.15     0.16       2000 1.00
```

Table 5: Outputs from LME and BHM analyses. The seed-based correlation results at 21 ROIs from 124 subjects were fitted with LME (using R package lme4) and BHM (using Stan with 4 chains and 1,000 iterations) separately, in which overall ToMI was an explanatory variable. Random effects under LME correspond to group/entity-level effects plus family specific parameters (variance σ^2 for residuals) under BHM, while fixed effects under LME correspond to population-level effects under BHM. The parameter estimates from the LME output (a) and the BHM output (b) are very similar, even though priors were injected into BHM. All $\hat{R}$ values under BHM were less than 1.1, indicating that all the 4 chains converged well. The effective sizes for the population- and group/entity-level effect of ToMI were 468 and 947, respectively, enough to warrant quantile accuracy in summarizing the posterior distributions.

Table 6: Comparisons of results between the conventional GLM and BHM[a]

result / ROI	ToMI effect		standard error		2.5%		5%		95%		97.5%	
	GLM	BHM	GLM	BHM	GLM	BHM	GLM	BHM	GLM	BHM	GLM	BHM
R PCC	0.025	0.018	0.010	0.006	0.005	0.008	0.008	0.009	0.041	0.028	0.045	0.030
R TPJp	0.018	0.025	0.009	0.007	0.000	0.012	0.003	0.014	0.034	0.036	0.037	0.038
R Insula	-0.004	0.002	0.006	0.006	-0.015	-0.010	-0.014	-0.008	0.006	0.011	0.007	0.013
L IPL	0.020	0.014	0.008	0.006	0.005	0.003	0.008	0.004	0.033	0.024	0.035	0.026
L SFG	0.011	0.008	0.008	0.006	-0.004	-0.003	-0.001	-0.001	0.024	0.017	0.027	0.019
R IFG (BA45)	-0.006	0.000	0.007	0.005	-0.021	-0.011	-0.019	-0.009	0.006	0.008	0.008	0.010
R IFG (BA9)	-0.002	0.002	0.005	0.005	-0.012	-0.009	-0.010	-0.007	0.006	0.011	0.008	0.012
L MTG	-0.001	0.004	0.009	0.005	-0.019	-0.007	-0.016	-0.005	0.013	0.013	0.016	0.015
L CG	-0.004	-0.003	0.007	0.005	-0.017	-0.014	-0.015	-0.011	0.007	0.006	0.009	0.008
L IFG	-0.002	0.000	0.005	0.005	-0.012	-0.011	-0.010	-0.009	0.007	0.009	0.009	0.011
ACC	0.002	0.002	0.007	0.005	-0.012	-0.008	-0.009	-0.006	0.014	0.011	0.016	0.013
SGC	0.006	0.004	0.006	0.005	-0.007	-0.007	-0.005	-0.005	0.016	0.013	0.018	0.014
PCC/PrC	0.017	0.012	0.009	0.005	-0.001	0.001	0.002	0.003	0.032	0.021	0.035	0.023
dmMPFC	0.014	0.010	0.009	0.005	-0.004	-0.001	-0.001	0.001	0.029	0.019	0.032	0.021
L TPJ	0.018	0.015	0.008	0.005	0.001	0.005	0.004	0.007	0.031	0.025	0.034	0.026
L vBG	0.001	0.003	0.005	0.005	-0.009	-0.008	-0.007	-0.006	0.010	0.011	0.012	0.012
R vBG	0.001	0.003	0.005	0.005	-0.009	-0.008	-0.007	-0.006	0.009	0.012	0.011	0.014
L aMTS/aMTG	0.022	0.013	0.009	0.006	0.005	0.003	0.007	0.005	0.036	0.023	0.039	0.025
R Amy/Hippo	-0.003	0.002	0.006	0.005	-0.014	-0.009	-0.012	-0.007	0.006	0.011	0.008	0.012
L Amy/Hippo	-0.004	0.001	0.006	0.005	-0.016	-0.010	-0.014	-0.008	0.005	0.010	0.007	0.012
vmPFC	0.015	0.009	0.008	0.006	-0.001	-0.001	0.002	0.000	0.029	0.019	0.031	0.021

[a] The pooling or shrinkage effect among the 21 ROIs under BHM can be seen for both ToMI effect (in the unit of z-score per one unit of behavior score) and its standard error (*cf.* Fig. 2) in the sense that they are dragged from the extremes to the center. The percentages show the percentile confidence intervals for the conventional GLM with no pooling and the HPD intervals for BHM, respectively. Rows in green indicate that the corresponding effect lies within the positive domain of the 95% HPD interval under BHM, revealing strong evidence for the behavior effect; rows in yellow indicate that the corresponding effect lies within the positive domain of the 90% HPD interval under BHM (or the 95% quantile interval if the effect sign is *a priori* known), revealing moderate evidence for the behavior effect. The conventional ROI-based GLM revealed a similar pattern but with different effect estimates and distributions due to the isolated treatment among the ROIs; however, none of the 21 ROIs would survive FWE correction under NHST. Unlike the popular practice of sharp thresholding under NHST, more customized quantile intervals (e.g., 10%, 50% and 90%), if desirable, can be added in the final reporting in order to make corresponding inferences.

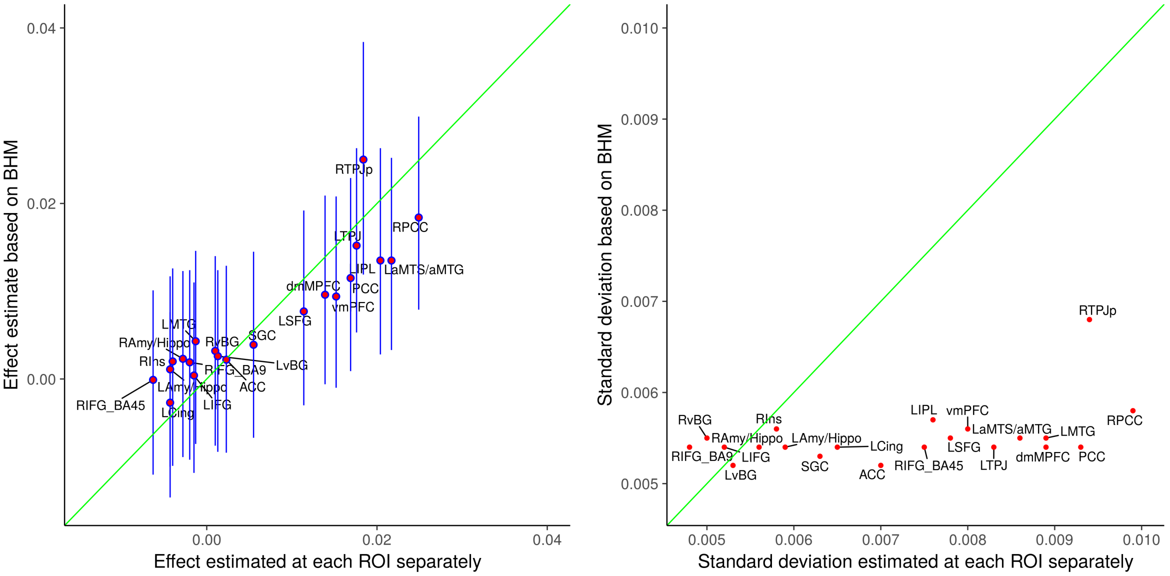
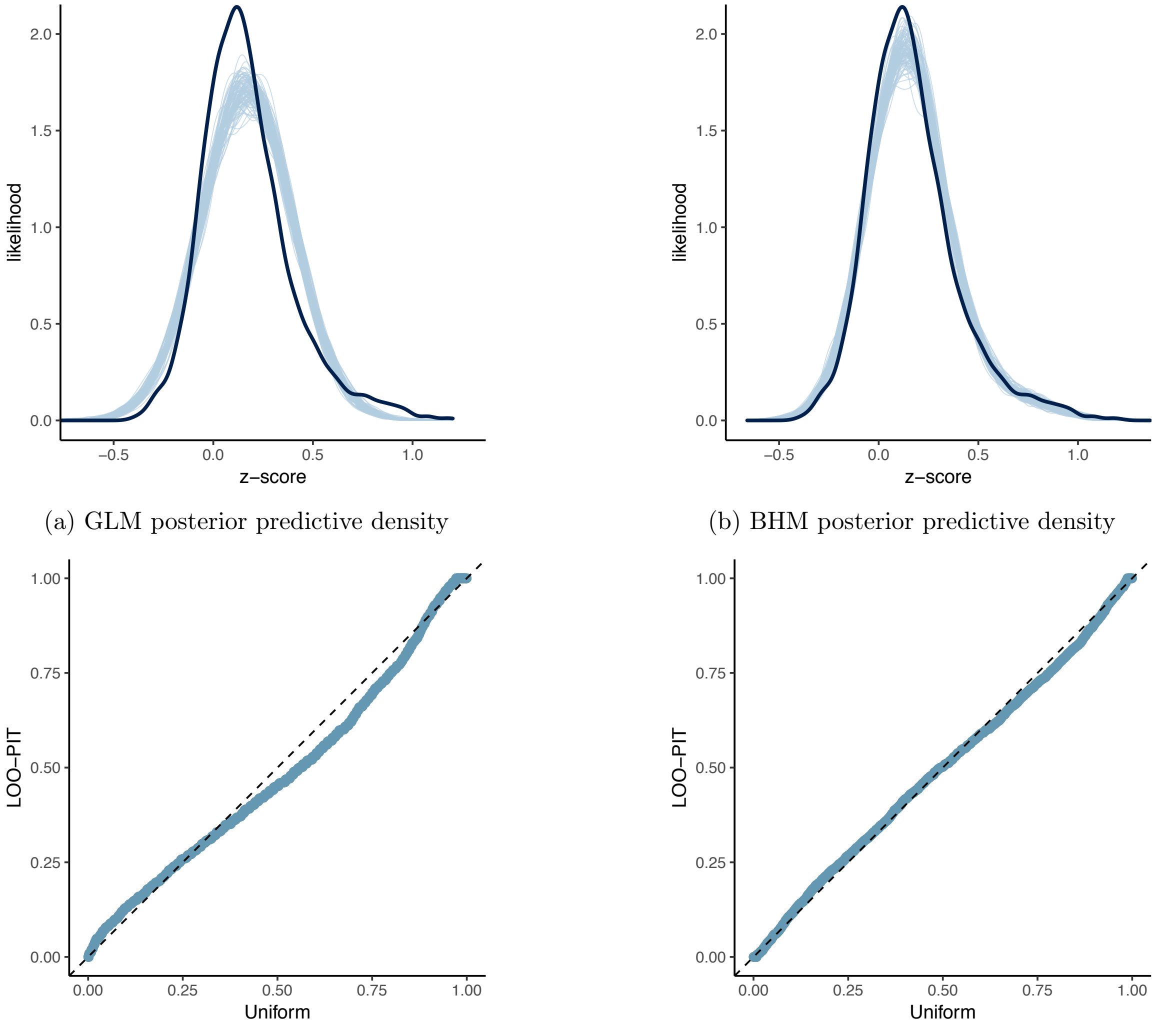


Figure 2: Illustration of shrinkage impact on effect of ToMI (left) and standard deviation (right) under BHM relative to univariate GLM (x axis). The diagonal line (green) serves as a reference under which no shrinkage occurs, and the blue bar for each effect estimate indicates the 95% posterior quantile interval. The shrinkage phenomenon reflects the fact that, under BHM, the effects are dragged toward the center among the ROIs except for the region of R TPJp that was actually pulled away from the center due to intertwined impact within the variance-covariance structures among the ROIs and subjects.



(a) GLM posterior predictive density

(b) BHM posterior predictive density

(c) GLM cross-validation: Q-Q plot (uniform)

(d) BHM cross-validation: Q-Q plot (uniform)

Figure 3: Model performance comparisons through posterior predictive check between conventional univariate GLM (a and c) and BHM (b and d). The subfigures a and b show the posterior predictive density overlaid with the raw data from the 124 subjects at the 21 ROIs for GLM and BHM, respectively: solid black curve is the raw data at the 21 ROIs with linear interpolation while the fat curve in light blue is composed of 100 sub-curves each of which corresponds to one draw from the posterior distribution based on the respective model. The differences between the two curves indicate how well the respective model fits the raw data. BHM fitted the data better than GLM at the peak and both tails as well as the skewness because pooling the data from both ends toward the center through shrinkage clearly validates our adoption of BHM. The subfigures c and d contrast GLM and BHM through cross-validation with leave-one-out log predictive densities through the calibration of marginal predictions from 100 draws; the calibration is assessed by comparing probability integral transformation (PIT) checks to the standard uniform distribution. The diagonal dished line indicates a perfect calibration: there are some suboptimal calibration for both models, but BHM is clearly a substantial improvement over GLM. To simulate the posterior predictive data for the conventional ROI-based approach (a and c), the original 21 GLMs were combined into one GLM so that the variances were pooled; in addition, the combined GLM was Bayesianized with a noninformative uniform prior for the population parameters (with no entity-level parameters in this case).

of 0.005 (Table 3) in the whole brain analysis were confirmed with the ROI-based BHM (Table 6). Moreover, another four regions (L aMTS/aMTG, R TPJp, vmPFC, dmPFC) revealed some evidence of ToMI effect under BHM. In contrast, these four regions did not stand out in the whole brain analysis after the application of FWE correction at the cluster level regardless of the voxel-wise p-threshold (Table 3), even though they would have been evident if the cluster size requirement were not as strictly imposed.

Discussion

Current approaches to correcting for FPR

In the conventional statistics framework, the thresholding bar ideally plays the role of winnowing the wheat (true effect) from the chaff (random noise), and a p-value of 0.05 is commonly adopted as a benchmark for comfort in most fields. However, one big problem facing the correction methods for multiple testing is the arbitrariness surrounding the thresholding, in addition to the arbitrariness of 0.05 itself. Both Monte Carlo simulations and random field theory start with a voxel-wise probability threshold (e.g., 0.01, 0.005, 0.001) at the voxel (or node) level, and a spatial threshold is determined in cluster size so that overall FPR can be properly controlled at the cluster level. If clusters are analogized as islands, each of them may be visible at a different sea level (voxel-wise p-value). As the cluster size based on statistical filtering plays a leveraging role, with a higher statistical threshold leading to a smaller cluster cutoff, a neurologically or anatomically small region can only gain ground with a low p-value while large regions with a relatively large p-value may fail to survive the criterion. Similarly, a lower statistical threshold (higher p) requires a higher cluster volume, so smaller regions have little chance of reaching the survival level. In addition, this arbitrariness in statistical threshold at the voxel level poses another challenge for the investigator: one may lose spatial specificity with a low statistical threshold since small regions that are contiguous may get swamped by the overlapping large spatial extent; on the other hand, sensitivity may have to be compromised for large regions with low statistic values when a high statistical threshold is chosen. A recent critique on the rigor of cluster formation through parametric modeling (Eklund et al., 2016) has resulted in a trend to require a higher statistical thresholding bar (e.g., with the voxel-wise threshold below 0.005 or even 0.001); however, the arbitrariness persists because this trend only shifts the probability threshold range.

As an alternative to parametric methods, an early version of permutation testing (Nichols and Holmes, 2001) starts with the construction of a null distribution through permutations in regard to a maximum statistic (either maximum testing statistic or maximum cluster size based on a predetermined threshold for the testing statistic). The original data is assessed against the null distribution, and the top winners at a designated rate (e.g., 5%) among the testing statistic values or clusters will be declared as the surviving ones. While the approach is effective in maintaining the nominal FWE level, two problems are embedded with the strategy. First of all, the spatial properties are not directly taken into consideration in the case of maximum testing statistic. For example, an extreme case to demonstrate the spatial extent issue is that a small cluster (or even a single voxel) might survive the permutation testing as long as its statistic value is high enough (e.g., $t(20) = 6.0$) while a large cluster with a relatively small maximum statistic value (e.g., $t(20) = 2.5$) would fail to pass the filtering. The second issue is the arbitrariness involved in the primary thresholding for the case of maximum cluster size: a different primary threshold may end up with a different set of clusters. That is, the reported results may likely depend strongly on an arbitrary threshold.

Addressing these two problems, a later version of permutation testing (Smith and Nichols, 2009) takes an integrative consideration between signal strength and spatial relatedness, and thus solves both problems involving the earlier version of permutation testing[4]. Such an approach has been implemented in programs such as Randomise and PALM in FSL using threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009) and in 3dttest++ in AFNI using equitable thresholding and clusterization (ETAC) (Cox, 2018). Nevertheless, the adoption of permutations in neuroimaging, regardless of the specific version, is not directly about the concern

[4] A single voxel is still possible, but much less likely, to survive this correction approach.

of distribution violation as in the classical nonparametric setting (in fact, a pseudo-t value is still computed at each voxel in the process); rather, it is the randomization among subjects in permutations that creates a null distribution against which the original data can be tested at the whole brain level.

Furthermore, all of those correction methods, parametric and nonparametric, are still meant to use spatial extent or the combination of spatial extent and signal strength as a filter to control the overall FPR at the whole brain level. They all share the same hallmark of sharp thresholding at a preset acceptance level (e.g., 5%) under NHST, and they all use spatial extent as a leverage, penalizing regions that are anatomically small and rewarding large smooth regions. Due to the unforgiving penalty of correction for multiple testing, some workaround solutions have been adopted by focusing the correction on a reduced domain instead of the whole brain. For example, the investigator may limit the correction domain to gray matter or regions based on previous studies. Putting the justification for these practices aside, accuracy is a challenge in defining such masks; in addition, spatial specificity remains a problem, shared by the typical whole brain analysis, although to a lesser extent.

Questionable practices under NHST

Univariate GLM may work reasonably well if the following two conditions can be met: 1) no multiple testing, and 2) high signal-to-noise ratio (strong effect and high precision measurement) as illustrated in the lower triangular part of Table 2. However, neither of the two conditions is likely satisfied with typical neuroimaging data. Due to the stringent requirements of correction for multiple testing across thousands of resolution elements in neuroimaging, a daunting challenge facing the community is the power inefficiency or high type II errors under NHST. Even if prior information is available as to which ROIs are potentially involved in the study, an ROI-based univariate GLM would still be obliged to share the burden of correction for multiplicity equally and agnostically to any such knowledge. The modeling approach usually does not have the luxury to survive the penalty, as shown with our experimental data in Table 6, unless only a few ROIs are included in the analysis.

Furthermore, with many low-hanging fruits with relatively strong signal strength (e.g., 0.5% signal change or above) having been largely plucked, the typical effect under investigation nowadays is usually subtle and likely small (e.g., in the neighborhood of 0.1% signal change). Compounded with the presence of substantial amount of noise and suboptimal modeling (e.g., ignoring the HDR shape subtleties), detection power is usually low. It might be counterintuitive, but one should realize that the noisier or more variable the data, the less one should be confident about any inferences based on statistical significance, as illustrated with the type S and type M errors in Figure 1. With fixed threshold correction approaches, small regions are hard to funnel through the FPR criterion even if their effect magnitude is the same as or even higher than those larger regions. Even for approaches that take into consideration both spatial extent and effect magnitude (TFCE, ETAC), small regions remain disadvantaged when their effect magnitude is at the same level as their larger counterparts.

Current knowledge regarding brain activations has not reached a point where one can make accurate dichotomous claims as to whether a specific brain region under a task or condition is activated or not; lack of underlying "ground truth" has made it difficult to validate any but the most basic models. The same issue can be raised about the binary decision as to whether the response difference under two conditions is either the same or different. Therefore, a pragmatic mission is to detect activated regions in terms of practical, instead of statistical, significance. The conventional NHST runs against the idealistically null point H_0, and declares a region having no effect based on statistical significance with a safeguard set for type I error. When the power is low, not only reproducibility will suffer, but also the chance of having an incorrect sign for a statistically significant effect be substantial (Table 2 and Fig. 1). Only when the power reaches 30% or above does the type S error rate become lpw. Publication bias due to the thresholding funnel contributes to type S and type M errors as well.

The sharp thresholding imposed by the widely adopted NHST strategy uses a single threshold through which a high-dimension dataset is funneled. An ongoing debate has been simmering for a few decades regarding the legitimacy of NHST, ranging from cautionary warning against misuses of NHST (Wasserstein and Lazar, 2016), to tightening the significance level from 0.05 to 0.005 (Benjamin et al., 2017), to totally abandoning NHST as a gatekeeper (McShare et al., 2017; Amrhein and Greenland, 2017). The poor controllability of type S and type M errors is tied up with widespread problems across many fields. It is not a common practice nor a requirement in neuroimaging to report the effect estimates; therefore, power analysis for a brain region under a task or condition is largely obscure and unavailable, let alone the assessment of type S and type M errors.

Relating the discussion to the neuroimaging context, the overall or global FPR is the probability of having data as extreme as or more extreme than the current result, under the assumption that the result was produced by some "random number generator," which is built into algorithms such as Monte Carlo simulations, random field theory, and randomly shuffled data as pseudo-noise in permutations. It boils down to the question: are the data truly pure noise (even though spatially correlated to some extent) in most brain regions? Since controlling for FPR hinges on the null hypothesis of no effect, p-value itself is a random variable that is a nonlinear continuous function of the data at hand, therefore it has a sampling distribution (e.g., uniform(0,1) distribution if the null model is true). In other words, it might be under-appreciated that even identical experiments would not necessarily replicate an effect that is dichotomized in the first one as statistically significant (Lazzeroni et al., 2016). The common practice of reporting only the statistically significant brain regions and comparing to those nonsignificant regions based on the imprimatur of statistic- or p-threshold can be misleading: the difference between a highly significant region and a nonsignificant region could simply be explained by pure chance. The binary emphasis on statistical significance unavoidably leads to an investigator only focusing on the significant regions and diverting attention away from the nonsignificant ones. More importantly, the traditional whole brain analysis usually leads to selectively report surviving clusters conditional on statistical significance through dichotomous thresholding, potentially inducing type M errors, biasing estimates with exaggerated effect magnitude, as illustrated in Fig. 1.

Rigor, stringency, and reproducibility are lifelines of science. We think that there is more room to improve the current practice of NHST and to avoid information waste and inefficiency. Because of low power in FMRI studies and the questionable practice of binarized decisions under NHST, a Bayesian approach combined with integrative modeling offers a platform to more accurately account for the data structure and to leverage the information across multiple levels.

What Bayesian modeling offers

Almost all statistics consumers (including the authors of this paper) were *a priori* trained within the conventional NHST paradigm, and their mindsets are usually familiar with and entrenched within the concept and narratives of p-value, type I error, and dichotomous interpretations of results. Out of the past shadows cast by the theoretical and computation hurdles, as well as the image of subjectivity, Bayesian methods have gradually emerged into the light. One direct benefit of Bayesian inference, compared to NHST, is its concise and straightforward interpretation, as illustrated in Table 7.

Table 7: Interpretation differences between NHST and Bayesian framework

	Probability p	**Effect Interval** $[L, U]$
NHST	If H_0 is true, the probability of having the current result or more extreme is p (based on what would have occurred under other possible datasets); e.g., $P(\|T(\boldsymbol{y})\| > t_c \| \text{easy} = \text{difficult}) = p$, where $T(\boldsymbol{y})$ is a statistic (e.g., Student's t) based on data $\boldsymbol{y}$ and t_c is a threshold.	If the study is exactly repeated an infinite number of times, the percentage of those confidence intervals will cover the true effect is $1-p$; e.g., $P(L \leq \text{easy} - \text{difficult} \leq U) = 1-p$, where "easy - difficult" is treated as being fixed while L and U are random.
Bayesian	The probability of having the *current* result being different from zero is p (given the dataset); e.g., $P(\text{easy} - \text{difficult} < L \text{ or easy} - \text{difficult} > U \| \boldsymbol{y}) = p$, where L and U are lower and upper bounds of the $(1-p)100\%$ quantile interval.	The probability that the effect falls in the predictive interval is $1-p$ (given the data); e.g., $P(L \leq \text{easy} - \text{difficult} \leq U \| \boldsymbol{y}) = 1-p$, where "easy - difficult" is considered random while L and U are known conditional on data $\boldsymbol{y}$.

Even though the NHST modeling strategy literally falsifies the straw man null hypothesis H_0, the real intention is to confirm the alternative (or research) hypothesis through rejecting H_0; in other words, the falsification of H_0 is only considered an intermediate step under NHST, and the ultimate goal is the confirmation of the intended hypothesis. In contrast, under the Bayesian paradigm, the investigator's hypothesis is directly placed under scrutiny through incorporating prior information, model checking and revision. Therefore, the Bayesian paradigm is more fundamentally aligned with the hypothetico-deductivism axis along the classic view of falsifiability or refutability by Karl Popper (Gelman and Shalizi, 2013).

Practically speaking, should we fully "trust" the effect estimate at each ROI or voxel at its face value? Some may argue that the effect estimate from the typical neuroimaging individual or population analysis has the desirable property of unbiasedness, as asymptotically promised by central limit theory. However, in reality the asymptotic property requires a very large sample size, a luxury difficult for most neuroimaging studies to reach. As the determination of a reasonable sample size depends on signal strength, brain region, and noise level, the typical sample size in neuroimaging tends to get overstretched, creating a hotbed for a low power situation and potentially high type S and type M errors.

Another fundamental issue with the conventional univariate modeling approach in neuroimaging is the two-step process: first, pretend that the voxels or nodes are independent with each other, and build a separate model for each spatial element; then, handle the multiple testing issue using spatial relatedness to only partially, not fully, recoup the efficiency loss. In addition to the conceptual novelty for those with little experience outside the NHST paradigm, BHM simplifies the traditional two-step workflow with one integrative model, fully shunning the multiple testing issue.

Bayesian modeling has traditionally been burdened with computational costs; however, this situation has recently been substantially ameliorated by the availability of multiple software tools such as Stan. Technical issues aside, one direct benefit of Bayesian modeling is its interpretational convenience (Table 7). Unlike the p-value under NHST, which represents the probability of obtaining the current data assuming that no effect exists anywhere, a posterior distribution for an effect under the Bayesian framework directly and explicitly shows the uncertainty of our current knowledge conditional on the current data and the prior model. From the Bayesian perspective, we should not put too much faith in point estimates. In fact, not only should we not fully trust the point estimates from GLM with no pooling, but also should we not rely too much on the centrality (median, mean) of the posterior distribution from a Bayesian model. Instead, the focus needs to be placed more on the uncertainty, which can be characterized by the quantile intervals of the posterior distribution if summarization is required. Specifically, BHM, as demonstrated here with the ROI data, is often more conservative in the sense that it does not "trust" the original effect estimates as much as GLM, as shown in Fig. 2; additionally, in doing so, it fits the data better than the ROI-based GLM (Fig. 3). Furthermore,, the original multiple testing issue under the massively univariate platform is deflected within one unified model: it is the type S, not type I, errors that are considered crucial and controlled under BHM. Even though the posterior inferences at the 95% HPD interval in our experimental data were similar to the statistically significant results at the 0.05 level under NHST, BHM in general is more statistically conservative than univariate GLM under NHST, as shown with the examples in Gelman et al. (2012).

BHM borrows information across ROIs to improve the quality of the individual estimates and posterior distributions with the assumption of similarity among the regions. From the NHST perspective, BHM can still commit type I errors, and its FPR could be higher under some circumstances than, for example, its GLM counterpart. Such type I errors may sound alarmingly serious; however, the situation is not as severe as its appearance for two reasons: 1) the concept of FPR and the associated model under NHST are framed with a null hypothesis, which is not considered pragmatically meaningful in the Bayesian perspective; and 2) in reality, inferences under BHM most likely have the same directionality as the true effect because type S errors are well controlled across the board under BHM (Gelman and Teulinckx, 2000). Just consider which of the following two scenarios is worse: (a) when power is low, the likelihood under the NHST to mistakenly infer that the BOLD response to easy condition is higher than difficult could be sizable (e.g., 30%), and (b) with the type S error rate controlled below, for example, 3.0%, the BHM might exaggerate the magnitude difference between difficult and easy conditions by, for example, 2 times. While not celebrating (b), we expect that most researchers would view (1) as more problematic.

One somewhat controversial aspect of Bayesian modeling is the adoption of a prior for each hyperparameter, and that the prior is meshed with the data and gets updated into the posterior distribution. Some may consider that an Achilles' heel of Bayesian modeling is its subjectivity with respect to prior selection. With no intention to get tangled in the epistemological roots or the philosophical debates of subjectivity versus objectivity, we simply divert the issue to the following suggestion (Gelman and Hennig, 2017): replacing the term "subjectivity" with "multiple perspectives and context dependence," and, in lieu of "objectivity", focusing on transparency, consensus, impartiality, and correspondence to observable reality. Therefore, our focus here is the pragmatic aspect of Bayesian modeling: with prior distributions, we can make inferences for each ROI under BHM, which cannot be achieved under LME.

Since noninformative priors are adopted for population effects, the only impact of prior information incorporated into BHM comes from the distributional assumptions about those entities such as subjects and ROIs, which basically play the role of regularization: when the amount of data is large, the weakly informative priors levy a negligible effect on the final inferences; on the other hand, when the data do not contain enough information for robust inferences, the weakly informative priors prevent the distributions from becoming unsupportively dispersive. Specifically, for simple models such as Student's t-test and GLM, Bayesian approach renders similar inferences if noninformative priors are assumed. On the surface a noninformative prior does not "artificially" inject much "subjective" information into the model, and should be preferred. In other words, it might be considered a desirable property from the NHST viewpoint, since noninformative priors are independent of the data. Because of this "objectivity" property, one may insist that noninformative priors should be used all the time. Counterintuitively, a noninformative prior may become so informative that it could cause unbounded damage (Gelman et al., 2017). If we analyze the r ROIs individually as in the r GLMs (19), the point estimate for each effect θ_j is considered stationary, and we would have to correct for multiple testing due to the fact that r separate models are fitted independently. Bonferroni correction would likely be too harsh, which is the major reason that ROI-based analysis is rarely adopted in neuroimaging except for effect verification or graphic visualization. On the other hand, the conventional approach with the r GLMs (19) is equivalent to the BHM (21) by *a priori* assuming an improper flat prior for θ_j with the cross-ROI variability $\tau^2 = \infty$; that is, each effect θ_j can be any value with equal likelihood within $(-\infty, \infty)$. In the case of BOLD response, it is not necessarly considered objective to adopt a noninformative priori such as uniform distribution when intentionally ignoring the prior knowledge. In fact, we do have the prior knowledge that the typical effect from a 3T scanner has the same scale and lies within, for example, $(-4, 4)$ in percent signal change; this commonality can be utilized to calibrate or regularize the noise, extreme values, or outliers due to pure chance or unaccounted-for confounding effects (Gelman et al., 2012), which is the rationale for our prior distribution assumption of Gaussianity for both subjects and ROIs. Even for an effect for a covariate (e.g., the associate between behavior and BOLD response), it would be far-fetched to assume that τ^2 has the equal chance between, for example, 0 and 10^{10}.

Another example of information waste under NHST is the following. Negative or zero variance can occur in an ANOVA model while zero variance may show up in LME. Such occurrences are usually due to the full reliance on the data or a parameter boundary, and such direct estimates are barely meaningful: an estimate of cross-subject variability $\lambda^2 = 0$ in (21) indicates that all subjects have absolutely identical effects. However, a Bayesian inference is a tug of war between data and priors, and therefore negative or zero variance inferences would not occur because those scenarios from the data are regularized by the priors, as previously shown in ICC computations that are regularized by a Gamma prior for the variance components (Chen et al., 2017c).

In general, when there is enough data, weakly informative priors are usually drowned out by the information from the data; on the other hand, when data are meager (e.g., with small or moderate sample size), such priors can play the role of regularization (Gelman et al., 2017) so that smoother and more stable inferences could be achieved than would be obtained with a flat prior. In addition, a weakly informative prior for BHM allows us to make reasonable inferences at the region level while model quality can be examined through tools such as posterior predictive check and LOO cross-validation. Therefore, if we do not want to waste such prior knowledge for an effect bounded within a range in the brain, the commonality shared by all the brain regions can be incorporated into the model through a weakly informative prior and the calibration of partial pooling among the ROIs, thus eliding the step of correcting for multiple testing under NHST.

Bayesian algorithms have usually been considered meticulous and time consuming to achieve convergence, making their adoption for wide applications difficult and impractical for hierarchical models even with a dataset of small or moderate size. However, the rapid development of Bayesian implementations in Stan over the past few years has changed the landscape. In particular, Stan adopts static HMC Samplers and its extension, NUTS, and it renders less autocorrelated and more effective draws for the posterior distributions, achieving quicker convergence than the traditional Bayesian algorithms such as Metropolis-Hastings, Gibbs sampling, and so on. With faster convergence and high efficiency, it is now feasible to perform full Bayesian inferences for BHM with datasets of moderate size.

Advantages of ROI-based BHM

Bayesian modeling has long been adopted in neuroimaging at the voxel or node level (see, e.g., recent developments such as Westfall et al., 2017; Eklund et al., 2017; Mejia et al., 2017); nevertheless, correction for FWE would still have to be imposed as part of the model or as an extra step. In the current context, we formulate the data generation mechanism for each dataset through a progressive triplet of models on a set of ROIs: GLM → LME → BHM. The strength of hierarchical modeling lies in its capability of stratifying the data in a hierarchical or multilevel structure so that complex dependency or correlation structures can be properly accounted for coherently within a single model. Specifically applicable in neuroimaging is a crossed or factorial layout between the list of ROIs and subjects as shown in the LME equation (10) and its Bayesian counterpart (11).

Our adoption of BHM, as illustrated with the demonstrative data analysis, indicates that BHM holds some promise for neuroimaging and offers the following advantages over traditional approaches:

1) Instead of separately correcting for multiple testing, BHM incorporates multiple testing as part of the model by assigning a prior distribution among the ROIs (i.e., treating ROIs as random effects under the LME paradigm). In doing so, multiple testing is handled by conservatively shrinking the original effect toward the center; that is, instead of leveraging cluster size or signal strength, BHM leverages the commonality among ROIs.

2) A statistically identified cluster through a whole brain analysis is not necessarily anatomically or functionally meaningful. In other words, a statistically identified cluster is not always aligned well with a brain region for diverse reasons such as "bleeding" effect due to contiguity among regions, and suboptimal alignment to the template space, as well as spatial blurring. In fact, a cluster may overlap multiple brain regions or subregions. In contrast, as long as a region can be *a priori* defined, its statistical inference under BHM is assessed by its signal strength relative to its peers, not by its spatial extent, providing an alternative to the whole brain analysis with more accurate spatial specificity.

3) When prior information about the directionality of an effect is available on some, but not all, regions (e.g., from previous studies in the literature), one may face the issue of performing two one-tailed t-tests at the same time in a blindfold fashion due to the limitation of the massively univariate approach. The ROI-based approach disentangles the complexity since the posterior inference for each ROI can be made separately.

4) It may be trite to cite the famous quote of "all models are wrong" by George E. P. Box. However, the reality in neoroimaging is that model quality check is substantially lacking. When prompted, one may recognize the potential problems and pitfalls of a model. However, when discussing and reporting results from the model, the investigator tends to treat the model as if it were always true and then discusses statistical inferences without realizing the implications or ramifications of the model that fits poorly or even conflicts with data. Building,

comparing, tuning and improving models is a daunting task with whole brain data due to the high computational cost and visualization inconvenience. In contrast, model quality check is an intrinsic part of Bayesian modeling process. The performance of each model and the room for improvement can be directly examined through graphical display as shown in Fig. 3.

5) A full results reporting is possible for all ROIs under BHM. The conventional NHST focuses on the point estimate of an effect supported with statistical evidence in the form of a p-value. In the same vein, typically the results from the whole brain analysis are displayed with sharp-thresholded maps or tables that only show the surviving clusters with statistic- or p-values. In contrast, as the focus under the Bayesian framework is on the predictive distribution, not the point estimate, of an effect, the totality of BHM results can be summarized in a table as shown in Table 6, listing the predictive intervals in various quantiles (e.g., 50%, 75%, and 95%), a luxury that whole brain analysis cannot provide.

6) To some extent, the ROI-based BHM approach can alleviate the arbitrariness involved in probability thresholding with the current FPR correction practices. Even though it allows the investigator to present the whole results for all regions, for example, in a table format, we do recognize that the investigator would prefer to focus their scientific discussion on some regions with strong posterior evidence. In general, with all effects, regardless of the strength of their statistical evidence, reported in totality, the decision of choosing which effects to discuss in a paper should be based on cost, benefit, and probabilities of all results (Gelman et al., 2014). Specifically for neuroimaging data analysis, the decision still does not have to be solely from the posterior distribution; instead, we suggest that the decision be hinged on the statistical evidence from the current data, combined with prior information from previous studies.

For example, one may still choose the 95% HPD interval as an equivalent benchmark to the conventional p-value of 0.05 when reporting the BHM results. However, those effects with, say, 90% quantile intervals excluding 0 can still be discussed with a careful and transparent description, which can be used as a reference for future studies to validate or refute; or, such effects can be reported if they have been shown in previous studies. Moreover, rather than a cherry-picking approach on reporting and discussing statistically significant clusters in whole brain analysis, we recommend a principled approach in results reporting in which the ROI-based results be reported in totality with a summary as shown in Table 6 and be discussed through transparency and soft, instead of sharp, thresholding. We believe that such a soft thresholding strategy is more healthy and wastes less information for a study that goes through a strenuous pipeline of experimental design, data collection, and analysis.

Limitations of ROI-based BHM and future directions

ROIs can be specified through several ways depending on the specific study or information available regarding the relevant regions. For example, one can find potential regions involved in a task or condition including resting state from the literature. Such regions are typically reported as the coordinates of a "peak" voxel (usually highest statistic value within a cluster), from which each region could be defined by centering a ball with a radius of, e.g., 6 mm in the brain volume (or by projecting an area on the surface). Regions can also be located through (typically coordinate-based) meta analysis with databases such as NeuroSynth (http://www.neurosynth.org) and BrainMap (http://www.brainmap.org), with tools such as brain_matrix (https://github.com/fredcallaway/brain_matrix), GingerALE (http://brainmap.org/ale), Sleuth (http://brainmap.org/sleuth), and Scribe (http://www.brainmap.org/scr that are associated with the database BrainMap. Anatomical atlases (e.g., http://surfer.nmr.mgh.harvard.edu, http://www.med.harvard.edu/aanlib) and functional parcellations (e.g., Schaefer et al., 2017) are another source of region definition. As a different strategy, by recruiting enough subjects, one could use half of the subjects to define ROIs, and the other half to perform ROI-based analysis; similarly, one could scan the same set of subjects longer, use the first portion of the data to define ROIs, and the rest to perform ROI-based analysis.

The limitations of the ROI-based BHM are as follows.

1) Just as the FWE correction on the massively univariate modeling results is sensitive to the size of the full domain in which it is levied (whole brain, gray matter, or a user-defined volume), so the results from BHM will

depend to some extent on the number of ROIs (or which) ones included. For a specific ROI j, changing the composition among the rest of ROIs (e.g., adding an extra ROI or replacing one ROI with another) may result in a different prior distribution (e.g, $\theta_j \sim N(\mu, \tau^2)$ in BHM (11)) and a different posterior distribution for θ_j. It merits noting that the regions should be selected from the current knowledge and relevancy of the effect under investigation. Another subtlety regarding ROI composition is as follows. When the original data are skewed as shown in our dataset (longer right tails in Figure 3), BHM performed much better than GLM in accommodating the skewness; nevertheless, as there are more positive values than negative ones in the data, the negative values under BHM would be shrunk more leftward than their positive counterparts, and may result in less favorable inferences for the negative values relative to their positive counterparts.

2) ROI data extraction involves averaging among voxels within the region. Averaging, as a spatial smoothing or low-pass filtering process, condenses, reduces or dilutes the information among the voxels (e.g., 30) within the region to one number, and loses any finer spatial structure within the ROI. In addition, the variability of extracted values across subjects and across ROIs could be different from the variability at the voxel level.

3) The whole brain analysis has the chance to discover new regions. In contrast, when not all regions or subregions currently can be accurately defined, or when no prior information is available to choose a region in the first place, the ROI-based approach may miss any potential regions if they are not included in the model.

4) ROI-based analysis is conditional on the availability and quality of the ROI definition. One challenge facing ROI definition is the inconsistency in the literature due to inaccuracies across different coordinate/template systems and publication bias. In addition, some extent of arbitrariness is embedded in ROI definition; for example, a uniform adoption of a fixed radius may not work well due to the heterogeneity of brain region sizes.

5) The exchangeability requirement of BHM assumes that no differential information is available across the ROIs in the model. Under some circumstances, ROIs can be expected to share some information and not fully independent, especially when they are anatomically contiguous or more functionally related than the other ROIs (e.g., homologous regions in opposite hemisphere). Ignoring spatial correlations, if substantially present, may lead to underestimated variability and inflated inferences. In the future we will explore the possibility of accounting for such a spatial correlation structure. A related question is whether we can apply the BHM strategy to the whole brain analysis (e.g., shrinking the effects among voxels). Although tempting, such an extension faces serious issues, such as daunting computational cost, serious violation of the exchangeability assumption, and loss of spatial specificity.

Conclusion

The prevalent adoption of dichotomous decision making under NHST runs against the continuous nature of most quantities under investigation, including neurological responses, which has been demonstrated to be problematic through type S and type M errors when the signal-to-noise ratio is low. The conventional correction for FWE in neuroimaging data analysis is viewed as a “desirable” standard procedure for whole brain analysis because the criterion comes under NHST. However, it is physiologically unfeasible to claim that there is absolutely no effect in most brain regions; therefore, we argue that setting the stage only to fight the straw man of no effect anywhere is not necessarily a powerful nor efficient inference strategy. Inference power is further reduced by FWE correction due to the inefficiency involved in the massively univariate modeling approach. As BOLD responses in the brain share the same scale and range, the ROI-based BHM approach proposed here allows the investigator to borrow strength and effectively regularize the distribution among the regions, and it can simultaneously achieve meaningful spatial specificity and detection efficiency. In addition, it can provide increasing transparency on model building, quality control, and results reporting, and offers a promising approach to addressing two multiplicity issues: multiple testing and double sidedness.

Acknowledgments

The research and writing of the paper were supported (GC, PAT, and RWC) by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA, and by the NIH grant R01HD079518A to TR and ER. Much of the modeling work here was inspired from Andrew Gelman's blog. We are indebted to Paul-Christian Bürkner and the Stan development team members Ben Goodrich, Daniel Simpson, Jonah Sol Gabry, Bob Carpenter, and Michael Betancourt for their help and technical support. The simulations were performed in R and the figures were generated with the R package ggplot2 (Wickham, 2009).

Appendix A. Derivation of posterior distribution for BHM (11)

We start with the BHM system (11) with a known sampling variance σ^2,

$$y_{ij}|\pi_i, \theta_j \sim \mathcal{N}(\pi_i + \theta_j, \sigma^2),\ i = 1, 2, ..., n,\ j = 1, 2, ..., r.$$

Conditional on θ_j and prior $\pi_i \sim N(0, \lambda^2)$, the variance for the sampling mean at the jth ROI, $\bar{y}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \theta_j + \frac{1}{n}\sum_{i=1}^{n} \pi_i + \frac{1}{n}\sum_{i=1}^{n} \epsilon_{ij}$, is $\frac{\lambda^2+\sigma^2}{n}$; that is,

$$\bar{y}_{\cdot j}|\theta_j, \lambda^2 \sim N(\theta_j, \frac{\lambda^2+\sigma^2}{n}),\ j = 1, 2, ..., r.$$

With priors $\pi_i \sim N(0, \lambda^2)$ and $\theta_j \sim N(\mu, \tau^2)$, we follow the same derivation as in the likelihood (6), and obtain the posterior distribution,

$$\theta_j|\mu, \tau, \lambda, \boldsymbol{y} \sim \mathcal{N}(\hat{\theta}_j, V),\ \text{where } \boldsymbol{y} = \{y_{ij}\}, \hat{\theta}_j = \frac{\frac{n}{\lambda^2+\sigma^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}},\ V = \frac{1}{\frac{n}{\lambda^2+\sigma^2} + \frac{1}{\tau^2}},\ j = 1, 2, .., r.$$

When the sampling variance σ^2 is unknown, we can solve the LME counterpart in (10),

$$y_{ij} = \mu + \pi_i + \xi_j + \epsilon_{ij},\ i = 1, 2, ..., n,\ j = 1, 2, ..., r.$$

We then plug the estimated variances $\hat{\lambda}^2$, $\hat{\tau}^2$ and $\hat{\sigma}^2$ into the above posterior distribution formulas, and obtain the posterior mean and variance through an approximate Bayesian approach.

References

Amrhein, V., Greenland, S., 2017. Remove, rather than redefine, statistical significance. Nature Human Behaviour 1:0224.

Bates, B., Maechler, M., Bolker, B. Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1):1-48.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., É, Johnson, V., 2017. Redefine statistical significance. Nature Hum. Behav. 1, 0189

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289-300.

Chen, G., Saad, Z.S., Nath, A.R., Beauchamp, M.S., Cox, R.W., 2012. FMRI Group Analysis Combining Effect Estimates and Their Variances. NeuroImage 60:747-765.

Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear Mixed-Effects Modeling Approach to FMRI Group Analysis. NeuroImage 73:176-190.

Chen, G., Saad, Z.S., Adleman, N.E., Leibenluft, E., Cox, R.W., 2015. Detecting the subtle shape differences in hemodynamic responses at the group level. Front. Neurosci., 26 October 2015.

Chen, G., Taylor, P.A., Shin, Y.W., Reynolds, R.C., Cox, R.W., 2017a. Untangling the Relatedness among Correlations, Part II: Inter-Subject Correlation Group Analysis through Linear Mixed-Effects Modeling. Neuroimage 147:825-840.

Chen, G., Taylor, P.A., Cox, R.W., 2017b. Is the statistic value all we should care about in neuroimaging? Neuroimage 147:952-959.

Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Leibenluft, E., Brotman, M.A., Cox, R.W., 2017c. Intraclass correlation: improved modeling approaches and applications for neuroimaging. Human Brain Mapping. Human Brain Mapping. DOI: 10.1002/hbm.23909

Cohen, J., 1994. The earth is round ($p < .05$). American Psychologist 49(12):997-1003.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29:162-173. http://afni.nimh.nih.gov.

Cox, R.W., Chen, G., Glen, D.R., Reynolds, R.C., Taylor, P.A., 2017. FMRI Clustering in AFNI: False-Positive Rates Redux. Brain Connect. 7(3):152-171.

Cox, R.W., 2018. Equitable Thresholding and Clustering. In preparation.

Cox, R.W., Taylor, P.A., 2017. Stability of Spatial Smoothness and Cluster-Size Threshold Estimates in FMRI using AFNI. https://arxiv.org/abs/1709.07471

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS 113(28):7900-7905.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med. 33:636-647.

Gelman, A., 2015. Statistics and the crisis of scientific replication. Significance 12(3):23-25.

Gelman, A. 2016. The problems with p-values are not just with p-values. The American Statistician, Online Discussion.

Gelman, A., Carlin, J., 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science 1-11.

Gelman, A., Hennig, C., 2016. Beyond subjective and objective in statistics. http://www.stat.columbia.edu/~gelman/research/unpublished/objectivityr3.pdf

Gelman, A., Hill, J. Yajima, M., 2012. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness 5:189-211.

Gelman, A., Loken, E., 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology 66:8-38.

Gelman, A., Simpson, D., Betancourt, M., 2017. The prior can generally only be understood in the context of the likelihood. arXiv:1708.07487

Gelman, A., Tuerlinckx, F., 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics15: 373-390.

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. PNAS 109 (14), 5487-5492.

Gonzalez-Castillo, J., Chen, G., Nichols, T., Cox, R.W., 2017. Bandettini, P.A., Variance Decomposition for Single-Subject task-based fMRI activity estimates across many sessions. NeuroImage 154:206-218.

Lazzeroni, L.C., Lu, Y., Belitskaya-Lévy, I., 2016. Solutions for quantifying P-value uncertainty and replication power. Nature Methods 13:107-110.

Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis, 100, 1989-2001.

Loken E., Gelman, A., 2017. Measurement error and the replication crisis. Science 355(6325):584-585.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2017. Abandon Statistical Significance. arXiv:1709.07588

Mejia, A., Yue, Y.R., Bolin, D., Lindren, F., Lindquist, M.A., 2017. A Bayesian General Linear Modeling Approach to Cortical Surface fMRI Data Analysis. arXiv:1706.00959

Mueller, K., Lepsien, J., Möller, H.E., Lohmann, G., Commentary: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Front Hum Neurosci 11:345.

Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp. 15(1):1-25.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ziad S. Saad, Richard C. Reynolds, Brenna Argall, Shruti Japee, Robert W. Cox, 2004. SUMA: An interface for surface-based intra- and inter-subject analysis with AFNI. Proc. 2004 IEEE International Symposium on Biomedical Imaging, 1510-1513.

Schaefer, A., Kong, R., Gordon, E.M., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T., 2017. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral Cortex. In press.

Simmons, J. P., Nelson, L. D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22:1359-1366.

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

Stan Development Team, 2017. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. http://mc-stan.org

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing transparency through a multiverse Analysis. Perspect Psychol Sci.11(5):702-712.

Vandekar, S.N., Satterthwaite, T.D., Rosen, A., Ciric, R., Roalf, D.R., Ruparel, K., Gur, R.C., Gur, R.E., Shinohara, R.E., 2017. Faster family-wise error control for neuroimaging with a parametric bootstrap. Biostatistics.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on *p*-values: context, process, and purpose. The American Statistician 70:2, 129-133.

Vehtari, A., Gelman, A. Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27(5):1413-1432.

Westfall, J., Nichols, T.E., Yarkoni, T., 2017. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research 1:23.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. Journal of Cerebral Blood Flow and Metabolism, 12:900-918.

Xiao, Y., Geng, F., Riggins, T., Chen, G., Redcay, E., 2017. Neural correlates of developing theory of mind competence in early childhood. Under review.