An Open Publishing Platform of
the Organization for Human Brain Mapping
APERTURE NEURO

Sources of Information Waste in Neuroimaging: Mishandling Structures, Thinking Dichotomously, and Over-Reducing Data

Gang Chen,[a,*] Paul A. Taylor,[a] Joel Stoddard,[b] Robert W. Cox,[a] Peter A. Bandettini,[c] and Luiz Pessoa[d]

[a]*Scientific and Statistical Computing Core, NIMH, National Institutes of Health, Bethesda, MD, USA*
[b]*Department of Psychiatry, University of Colorado, Aurora, CO, USA*
[c]*Section on Functional Imaging Methods, NIMH, National Institutes of Health, Bethesda, MD, USA*
[d]*Department of Psychology, Department of Electrical and Computer Engineering, and Maryland Neuroimaging Center, University of Maryland, College Park, MD, USA*

ABSTRACT

Neuroimaging relies on separate statistical inferences at tens of thousands of spatial locations. Such massively univariate analysis typically requires an adjustment for multiple testing in an attempt to maintain the family-wise error rate at a nominal level of 5%. First, we examine three sources of substantial information loss that are associated with the common practice under the massively univariate framework: (a) the hierarchical data structures (spatial units and trials) are not well maintained in the modeling process; (b) the adjustment for multiple testing leads to an artificial step of strict thresholding; (c) information is excessively reduced during both modeling and result reporting. These sources of information loss have far-reaching impacts on result interpretability as well as reproducibility in neuroimaging. Second, to improve inference efficiency, predictive accuracy, and generalizability, we propose a Bayesian multilevel modeling framework that closely characterizes the data hierarchies across spatial units and experimental trials. Rather than analyzing the data in a way that first creates multiplicity and then resorts to a post hoc solution to address them, we suggest directly incorporating the cross-space information into one single model under the Bayesian framework (so there is no multiplicity issue). Third, regardless of the modeling framework one adopts, we make four actionable suggestions to alleviate information waste and to improve reproducibility: (1) model data hierarchies, (2) quantify effects, (3) abandon strict dichotomization, and (4) report full results. We provide examples for all of these points using both demo and real studies, including the recent Neuroimaging Analysis Replication and Prediction Study (NARPS).

Keywords: Bayesian multilevel modeling; data hierarchy; dichotomization; effect magnitude; information waste; multiple testing problem; result reporting

Correspondence: Gang Chen, Scientific and Statistical Computing Core, NIMH, National Institutes of Health, Bethesda, MD, USA. Email: gangchen@mail.nih.gov

Received: May 12, 2021

Accepted: November 08, 2021

DOI: 10.52294/ApertureNeuro.2022.2.ZRJI8542

INTRODUCTION

> *Statisticians classically asked the wrong question – and were willing to answer with a lie. They asked "Are the effects of A and B different?" and they were willing to answer "no."*
>
> *All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking "are the effects different?" is foolish.*
>
> John W. Tukey, "The Philosophy of Multiple Comparisons," Statistical Science (1991)

Functional magnetic resonance imaging (FMRI) is a mainstay technique of human neuroscience, which allows the study of the neural correlates of many functions, including perception, emotion, and cognition. The basic spatial unit of FMRI data is a *voxel* ranging from 1 to 3 mm on each side. As data are collected across time when a participant performs tasks or remains at "rest," FMRI datasets contain a time series at each voxel. Typically, tens of thousands of voxels are analyzed simultaneously. Such a "divide and conquer" approach through *massively univariate analysis* necessitates some form of multiple testing adjustment via procedures based on Bonferroni's inequality or false discovery rate.

Conventional neuroimaging inferences follow the null hypothesis significance testing framework, where the decision procedure dichotomizes the available evidence into two categories at the end. Thus, one part of the evidence survives an adjusted threshold at the whole-brain level and is considered *statistically significant* (informally interpreted as a "true" effect); the other part is ignored (often misinterpreted as "not true") and by convention omitted and hidden from public view (i.e., the file drawer problem).

A recent study (1) (referred to as NARPS hereafter) offers a salient opportunity for the neuroimaging community to reflect about common practices in statistical modeling and the communication of study findings. The study recruited 70 teams charged with the task of analyzing a particular FMRI dataset and reporting results; the teams simply were asked to follow data analyses routinely employed in their labs at the whole-brain voxel level (but note that nine specific research hypotheses were restricted to only three brain regions). NARPS found large variability in reported decisions, which were deemed to be sensitive to analysis choices ranging from preprocessing steps (e.g., spatial smoothing, head motion correction) to the specific approach used to handle multiple testing. Based on these findings, NARPS outlined potential recommendations for the field of neuroimaging research.

Despite useful lessons revealed by the NARPS investigation, the project also exemplifies the common approach in neuroimaging of generating categorical inferential conclusions as encapsulated by the "significant versus nonsignificant" maxim. In this context, we address the following questions:

1. Are conventional multiple testing adjustment methods informationally wasteful?
2. NARPS suggested that there was "substantial variability" in reported results across teams of investigators analyzing the same dataset. Is this conclusion dependent, at least in part, on the common practice of ignoring spatial hierarchy at the global level and drawing inferences binarily (i.e., "significant" vs. "nonsignificant")?
3. What changes can the neuroimaging field make in modeling and result reporting to improve reproducibility?

In this context, we consider inferential procedures not strictly couched in the standard null hypothesis significance testing framework. Rather, we suggest that multilevel models, particularly when constructed within a Bayesian framework, provide powerful tools for the analysis of neuroimaging studies given the data's inherent hierarchical structure. As our paper focuses on hierarchical modeling and dichotomous thinking in neuroimaging, we do not discuss the broader literature on Bayesian methods applied to FMRI (2).

MASSIVELY UNIVARIATE ANALYSIS AND MULTIPLE TESTING

We start with a brief refresher of the conventional statistical framework typically adopted in neuroimaging. Statistical testing begins by accepting the null hypothesis but then rejecting it in favor of the alternative hypothesis if the current data for the effect of interest (e.g., task A vs. task B) or potentially more extreme observations are unlikely to occur under the assumption that the effect is absolutely zero. Because the basic data unit is the voxel, one faces the problem of performing tens of thousands of inferences across space *simultaneously*. As the spatial units are not independent of one another, adopting an adjustment such as Bonferroni's is unreasonably conservative. Instead, the field has gradually settled into employing a cluster-based approach: what is the size of a spatial cluster that would be unlikely to occur under the null scenario?

Accordingly, a two-step procedure is utilized: first threshold the voxelwise statistical evidence at a particular (or a range of) voxelwise p-value (e.g., 0.001) and then consider only contiguous clusters of evidence (Fig. 1). Several adjustment methods have been developed to address multiple testing by leveraging the spatial relatedness among neighboring voxels. The stringency of the procedures has been extensively debated over the past decades, with the overall probability of having clusters of a minimum spatial extent given a null effect estimated by two common approaches: a parametric method (3,4) and a permutation-based adjustment (5). For the former, recent recommendations have resulted in the convention of adopting a primary threshold of voxelwise $p = 0.001$ followed by cluster size determination (6,7); for the latter, the threshold is based on the integration between a range of statistical evidence and the associated spatial extent (5).

Problems of multiple testing adjustments

At least five limitations are associated with multiple testing adjustments leveraged through spatial extent (8).

1. *Conceptual inconsistency*. Consider that the staples of neuroimaging research are the maps of statistical evidence and associated tables. Both typically present only the statistic (e.g., t) values. However, this change of focus is inconsistent with cluster-based inference: after multiple testing adjustment, the proper unit of inference is the cluster, not the voxel. Once "significant" clusters are determined, one *should* only speak of clusters and the voxels inside each cluster *should* no longer be considered meaningful inferentially. In other words, the statistical evidence for each surviving cluster is deemed at the "significance" level of 0.05 and the voxelwise statistic values

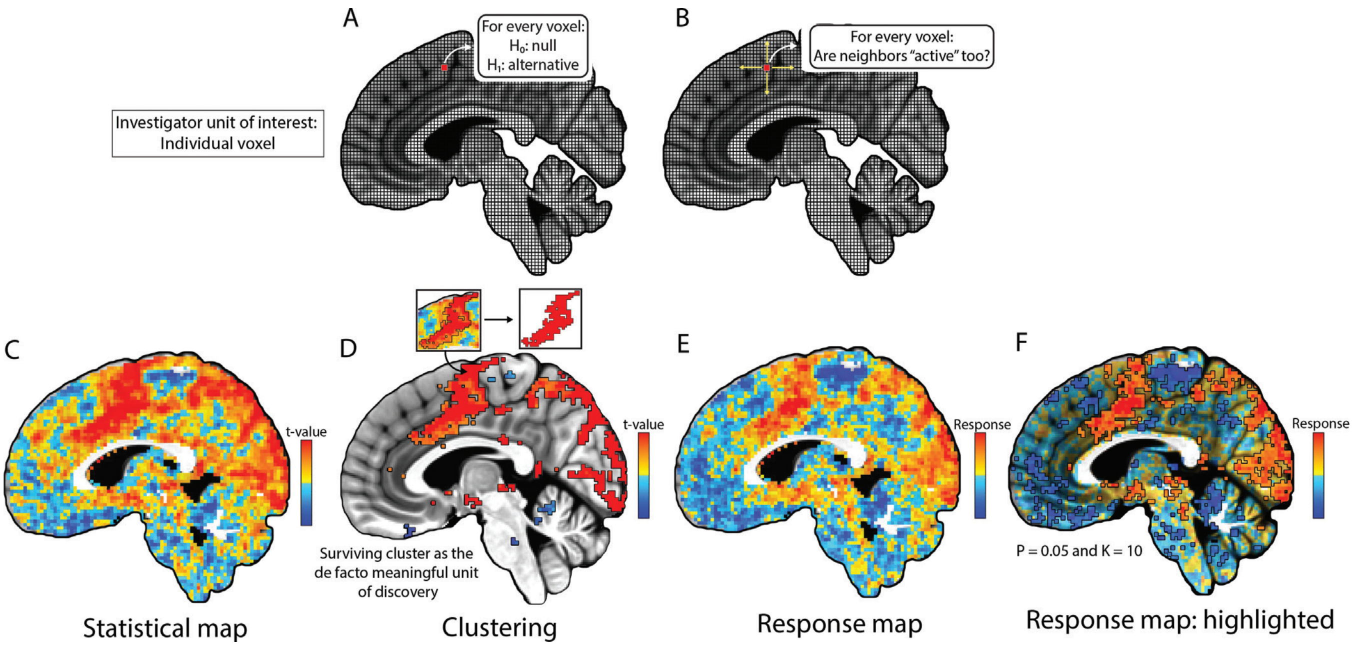


Fig. 1. Statistical inferences in neuroimaging. (A) Schematic view of standard analysis: each voxel among tens of thousands of voxels is tested against the null hypothesis (voxel not drawn to scale). (B) Clusters of contiguous voxels with strong statistical evidence are adopted to address the multiple testing problem. (C) Full statistical evidence for an example dataset is shown without thresholding. (D) The statistical evidence in (C) is thresholded at voxelwise p = 0.001 and a cluster threshold of 20 voxels. The left inset shows the voxelwise statistical values from (C) while the right inset illustrates the surviving cluster. (E) The map of effect estimates that complements the statistical values in (C), providing percent signal change or other index of response strength, is shown. (F) For presenting results, we recommend showing the map of effect estimates, while using the statistical information for little or moderate thresholding (e.g., cluster threshold K = 20 voxels at voxelwise p = 0.05): "highlight" parts with strong statistical evidence, but do not "hide" the rest.

lose direct interpretability. Therefore, voxel-level statistic values in brain maps and tables in the literature should not be taken at face value.

2. *Spatial ambiguity*. As a cluster is purely defined through statistical evidence, it is usually not aligned with any anatomical region, presenting a spatial specificity problem. To resolve the issue, the investigator typically reduces the cluster to a "peak" voxel with the highest statistical value and uses its location as evidence for the underlying region. A conceptual inconsistency results from these two transitional steps: one from a cluster to its peak voxel and then another from the voxel to an anatomical region. Furthermore, when a cluster spans over more than one anatomical region, no definite solutions are available to resolve the inferential difficulty. Although these issues of conceptual inconsistency and spatial ambiguity have been discussed in the past (7,8), it remains underappreciated, and researchers commonly do not adjust their presentations to match the cluster-level effective resolution.

3. *Heavy penalty against small regions.* With the statistical threshold at the spatial unit level traded off with cluster extent, larger regions might be able to survive with relatively weaker statistical evidence while smaller regions would have to require much stronger statistical strength. Therefore, multiple testing adjustments always penalize small clusters. Regardless of the specific adjustment method, anatomically small regions (e.g., those in the subcortex) are intrinsically disadvantaged even if they have the same amount of statistical evidence. In other words, ideally, the evidence for a brain region should be assessed solely in light of its effect magnitude, not dependent on its anatomical size; thus, the conventional multiple testing adjustment approaches are unfair to small regions because of its heavy reliance on spatial relatedness among the contiguous neighborhood.

4. *Sensitivity to data domain.* As the penalty for multiplicity becomes heavier when more spatial units are involved, one could explore various surviving clusters by changing the data space, resulting in some extent of arbitrariness: even though the data remain the same, a cluster may survive or fail depending on the investigator's choice of spatial extent for the data. Because of this vulnerability, it is not easy to draw a clear line between a justifiable reduction of data and an exploratory search (e.g., "small volume correction").

5. *Difficulty of assigning uncertainty.* As the final results are inferred at the cluster level, there is no clear uncertainty that can be attached to the effect magnitude at the cluster level. On the one hand, a cluster either survives or not under a dichotomous decision. On the other hand, due to the interpretation difficulty of voxel-level statistical evidence,

it remains challenging to have, for example, a standard error ("error bar") associated with the average effect at the cluster level.

Multiple testing adjustments are tied to the current result reporting practice. It is worth remembering a key goal of data processing and statistical modeling: to take a massive amount of data that is not interpretable in its raw state and to extract and distill meaningful information. The preprocessing parts aim to reduce distortion effects, whereas statistical models intend to account for various effects. Overall, there is a broad trade-off along the "analysis pipeline": we increase the digestibility of the information at the cost of reducing information. Fig. 2A illustrates these key aspects of the process of information extraction in standard FMRI analysis. The input data of time series across the brain for multiple participants are rich in information but of course not easily interpretable or "digestible." After multiple preprocessing steps followed by massively univariate analysis, the original data are condensed into two pieces of information at each spatial unit: the point estimate and the standard error. Whereas this process entails considerable reduction of information, it produces usefully digestible results; we highlight this trade-off in Fig. 2B. Here, "information" refers broadly to the amount and content of data present in a stage (e.g., for the raw data, the number of groups, participants, time series lengths). "Digestibility" refers to the ease with which the data are presentable and understandable (e.g., two 3D volumes vs. one; a 3D volume vs. a table of values). Following common practice, many investigators discard effect magnitude information to focus on summary statistics, which are then used to make binarized inferences by taking into account multiple testing. These steps certainly aid in reporting results and summarizing potentially some notable aspects of the data. However, we argue that the overall procedure leads to information waste and that the gained digestibility is relatively small (in addition to generating problems when results are compared across studies). Whereas we focus our discussion on whole-brain voxel-based analyses, similar issues apply in other types of analysis for region-based and matrix-based data.

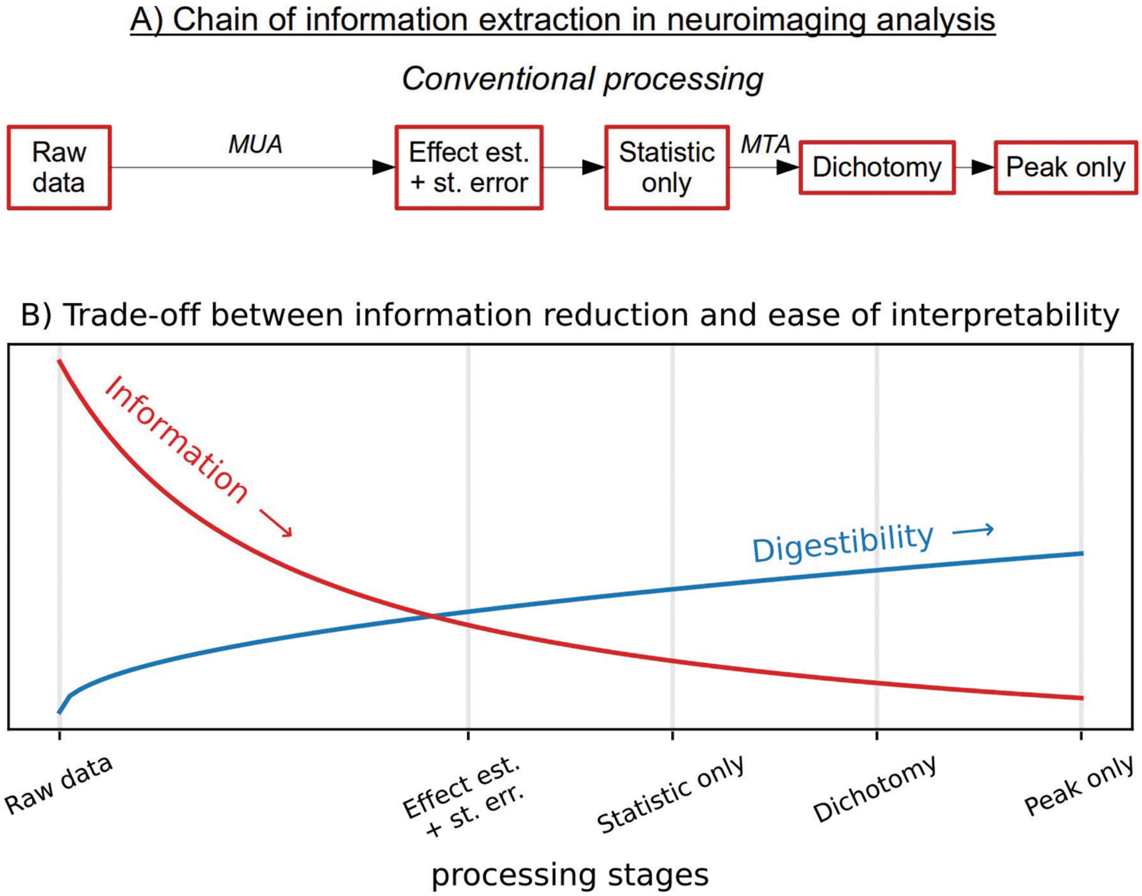


Fig. 2. **A schematic of conventional information extraction in neuroimaging.** (A) The processing chain starts with raw data. Massively univariate analysis (MUA) produces a point estimate and its uncertainty (standard error) at every spatial unit. These are reduced to a single statistic map, which is then dichotomized using thresholding through multiple testing adjustment (MTA); finally, the analyst summarizes the regions based solely on their peak values, ignoring spatial extent. (B) The inherent trade-off between "information" and "digestibility" (*y*-axis has arbitrary units). While summarizing peak locations of dichotomized regions is a highly digestible form of output, this also entails a severe information loss. Here, we argue that providing effect estimates and standard errors, if possible, would be preferable, striking a better balance between information loss and interpretability.

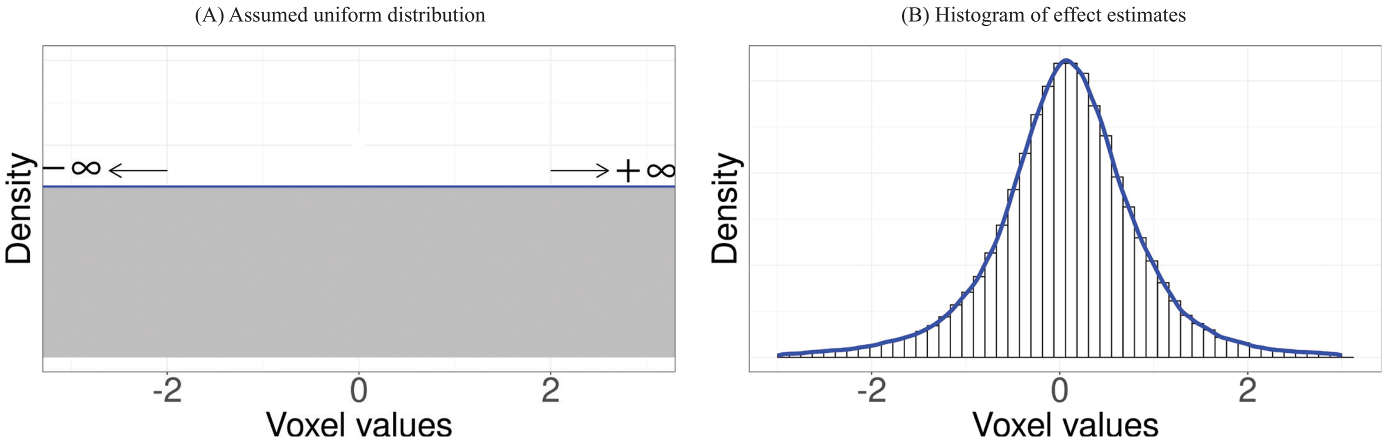


Fig. 3. Distributions of effects ("activation strength" in percent signal change) across space. (A) In massively univariate analysis, effects across all spatial units (voxels) are implicitly assumed to be drawn from a uniform distribution. Accordingly, the effect at each spatial unit can assume any value within (−∞, +∞) with equal likelihood. (B) Histogram of effect estimates (percent signal change) across 153,768 voxels in the brain from a particular study. Contrary to the assumption of uniform distribution implicitly made in massively univariate models, the effects approximately trace a Gaussian (or Student's *t*) distribution.

The implicit assumption of massively univariate analysis

Massively univariate analysis, by definition, models all voxels simultaneously with the assumption that all voxels (typically covering the entire brain) are unrelated to one another and that they do not share information. As a corollary, this also assumes that all possible effects have the same probability of being observed, which is to say that the effects follow a uniform distribution from −∞ to +∞ (Fig. 3A), at times discussed as the *principle of indifference* (9) or the *principle of insufficient reason* (10). Adopting this "indifference" approach might be reasonable, especially when the distribution of effects is unknown. However, it may result in information loss and lead to costly statistical accommodations.

In this context, we ask the following question: Do FMRI effects across the brain actually follow a uniform distribution, as tacitly assumed in massively univariate analysis, or are they closer to a symmetric bell-shaped distribution? We suggest that a better starting point would be a Gaussian (or possibly something with heavier tails, like Student's *t*) distribution (Fig. 3B). Conceptually, a Gaussian distribution is a reasonable choice if the effects track an average while also exhibiting a certain extent of variability.

Two direct consequences of massively univariate analysis are information waste and overfitting. Under the principle of insufficient reason, one trusts the *local* "unbiased" point estimates while "correcting" the extent of statistical evidence among neighboring spatial units during multiple testing adjustment; however, the loss of modeling efficiency and accuracy at the global level can only be partly recouped at the neighborhood, not *global*, level (8). In addition to potentially excessive penalties due to information waste, the principle of indifference has another important ramification: *overfitting*. As spatial units are treated as parallel entities – not part of the data hierarchy – in the model, global information shared across space is not leveraged and calibrated, leading to the loss of modeling efficiency. In other words, under massively univariate analysis, the model is free to fit the voxel's data in any way it can as all possible effect magnitudes are equally likely. As the field of machine learning has demonstrated repeatedly, overfitting is a serious problem because of compromised generalizability: Is it possible to learn from a sample to predict out-of-sample test cases? Thus, whereas the massively univariate approach offers unbiased estimates at the spatial unit level (via least squares or maximum likelihood), it tends to fit individual voxels overly close to the sample data at hand. Consequently, it may lead to a suboptimal trade-off between bias and variance and pay the cost of overfitting the data with reduced predictive accuracy when future data are considered.

What can be done to address the issues of information waste and overfitting? As a first step, we suggest that voxelwise modeling should take a holistic view, considering the effects as distributed normally (or according to Student's *t*). The reasoning here is analogous to when we assume that effects are normally distributed *across subjects* (or "random-effects" in linear mixed-effects modeling) in neuroimaging studies, allowing inferences at the population level. In a similar fashion, we propose conceptualizing voxel-level effects in terms of sampling from a normally distributed hypothetical pool of effects, instead of adopting the stance of complete ignorance (i.e., uniform distribution).

Technically, we can say that the effect distribution across spatial units, $N(\mu, \sigma^2)$, forms a *prior distribution* in the Bayesian sense where the two hyperparameters, the mean μ and the standard deviation σ, are basically estimated from the data. On the one hand, the variability of the data across spatial units (see Fig. 3B) determines the magnitude of σ. On the other hand, the estimated σ influences the estimates of μ across the spatial units through a process of "information sharing," regularization, or *partial pooling*. For example, if most of the individual effects across space are estimated to be small and close to zero, σ is estimated to be small, which further tends to decrease the individual effects, a situation also referred to as *shrinkage*.

We do not claim that the conventional approach is not valid. Instead, we suggest that the indifference assumption is an inefficient way of modeling the data, which can benefit from information sharing across space. Note that when NARPS summarized team results to make meta-analytic statements, they did not assume a uniform distribution of effects across teams; instead, they assumed that the results across studies would follow a Gaussian distribution. In other words, they did not treat the teams as "isolated trees." Interestingly, they did not adjust for multiple testing when interpreting individual team inferences, even though those 70 teams simultaneously analyzed the data and provided separate results. We agree that the adoption of a Gaussian prior is a sensible approach in their meta-analyses: it assumes that the results track an average population effect while exhibiting variability across teams. However, we propose that such utilization of priors does not have to be limited to or stopped at meta-analyses across different analytical pipelines; rather, information integration through a "forest perspective" can be equally applied to modeling across all hierarchies, including the levels of voxels, regions, experimental trials, and participants.

PROBLEMS OF DICHOTOMOUS THINKING

Data compression is essential in science so that complex information originating from large datasets can be encapsulated in terms of key findings (Fig. 2). Nevertheless, we believe that neuroimaging's common practice of adhering to multiple testing adjustments together with dichotomization ("significant or not") is detrimental to scientific progress. Take the process of examining the results by first insisting on the use of a cluster-based approach through a strict voxelwise threshold ($p < 0.001$) coupled with a minimum cluster extent (say, 50 voxels). In many instances, the analyst will miss the opportunity to make important novel observations; maybe some non-surviving clusters are just over 30 voxels (not to mention 49 voxels), for instance. The permutation-based approach to handling multiplicity suffers from the same issue.

In the last decade, statisticians and practitioners have extensively discussed pervasive issues with the practice of significance testing (11). As typically practiced in neuroimaging, solely focusing on and reporting statistical results that have survived significance filtering leads to issues such as overestimation ("winner's curse," publication bias (12,13), or type M error[14]) and type S error (incorrect sign) (14). A widespread problem is the disconnect between null hypothesis significance testing and the way investigators think of their research hypothesis. The *p*-value is the probability (or the extent of inconsistency or "surprise") of a random process generating the current data or *potentially more extreme observations* if a null effect were actually true (conditioned on the experimental design, the adopted model, and underlying assumptions). In contrast, an investigator is likely more interested in the probability of a research hypothesis (e.g., a positive effect) given the data. Misinterpretations of the *p*-value frequently lead to conceptual confusions (15). The *p*-values are also affected by the extent to which the model in question and its assumptions are suited for the data at hand.

Recognizing deep and entrenched research practices, the American Statistical Association has issued guidelines and proposed potential reforms (16). In our view, this important debate has not penetrated the neuroimaging community sufficiently. Given the expense and risk of collecting FMRI data, it is important to embrace methods that address problems with "significance testing" while simultaneously decreasing information waste. In a nutshell, we believe experimental science and discovery is a highly complex process that cannot be simplified and reduced to drawing a sharp line with the use of thresholding procedures, regardless of their numerical stringency and formal mathematical properties.

Problems with boiling down complicated study designs into binary decisions are further aggravated by the empirical observation that, as discussed, effects across the brain tend to follow a Gaussian distribution (Fig. 3B). Consistent with this notion, one study reported that over 95% of the brain was engaged in a simple visual stimulation plus attention task when large trial samples were adopted on three subjects (17). In contrast, most studies in neuroimaging only report a few brain regions that happen to survive the artificial dichotomization based on the currently accepted spatial adjustment criteria. The large gap between the engagement in most regions and the few regions reported in the literature is likely due to limited sample sizes as well as to information waste. More generally, many domains of research appear to be characterized by a very large number of "small effects" as opposed to a few "large effects," including genetics (18,19) and most likely brain research itself. Thus, a data analysis framework, such as null hypothesis significance testing, that seeks to binarize results only using statistical evidence (while ignoring separate effect estimates and uncertainties) is potentially problematic. We conjecture that this could represent the case in neuroimaging, where effects are present across large numbers of spatial units (voxels or brain regions) at varying strengths. In addition, it is worth noting that, even though family-wise error rate is the major leverage adopted to control multiplicity in neuroimaging, the arbitrariness issue involved in dichotomization equally applies to other notions such as false discovery rate.

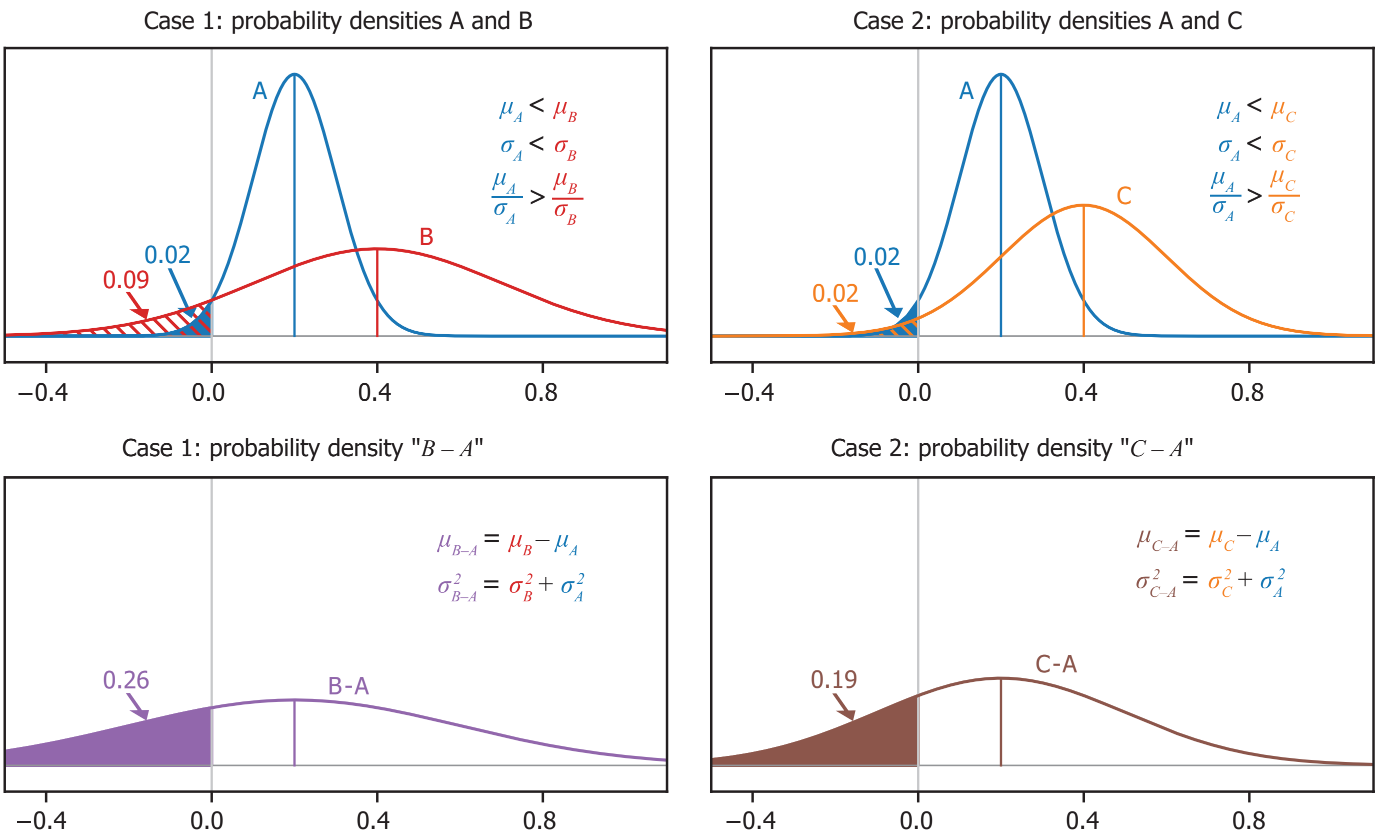


Fig. 4. Implications of dichotomization in conventional statistical practice. Case 1. What is the difference between a statistically significant result and one that does not cross a nominal threshold? Between the two hypothetical effects *A* and *B* that independently follow $N(\mu, \sigma^2)$ (upper left: $\mu_A = 0.2$, $\sigma_A = 0.1$ (blue); $\mu_B = 0.4$, $\sigma_B = 0.3$ (red)), only A would be considered statistically significant. As the difference between the two random variables associated with *A* and *B* follows $N(\mu_B - \mu_A, \sigma^2_B + \sigma^2_A) = N(0.2, 0.1)$, it is not considered statistically significant (lower left: $p = 0.26$, area under the density of $N(0.2, 0.1)$ on the left side of the gray line $x = 0$), and effect *B* is mostly larger than *A* with a probability of 0.74 (lower left: area under the density of $N(0.2, 0.1)$ on the right side of the gray line $x = 0$). **Case 2.** How much information is lost due to the focus on binary statistical decisions? The two hypothetical effects *A* and *C* that independently follow $N(\mu, \sigma^2)$ (upper right: $\mu_A = 0.2$, $\sigma_A = 0.1$ (blue); $\mu_C = 0.4$, $\sigma_C = 0.2$ (orange)) have the same *p*-values and would be deemed indistinguishable in terms of statistical evidence alone. However, as the difference between the two random variables associated with *A* and *C* follows $N(\mu_C - \mu_A, \sigma^2_C + \sigma^2_A) = N(0.2, 0.05)$, *C* is mostly larger than *A* with a probability of 0.81 (lower right: area under the density of $N(0.2, 0.05)$ on the right side of the gray line $x = 0$). This comparison illustrates the information loss when the sole focus is on statistics or *p*-value, which is further illustrated between the second and third blocks in Fig. 2.

We propose that a more productive approach is to refocus research objectives away from trying to uncover "real" effects. Specifically, more emphasis can be placed on discussing effects with stronger evidence, comparing large against small ones, or effects with smaller uncertainty against ones with larger uncertainty (Fig. 4, Case 2). Accordingly, methodological research goals should concentrate on developing an efficient experimental design and improving statistical modeling. More broadly, we advocate for approaches that are more accepting of the statistical uncertainty associated with data analysis, that is, more cognizant of inherent variability in data. In particular, investigators should not treat results that survive a particular threshold as "real" with the rest as "non-effects" and thus should not describe effects that survive as "facts." In other words, we recommend that one should avoid a result description with definitive certainty (e.g., null effect); even the typical language of "active/activated voxels/regions" comes with substantial perils. In general, we encourage further discussion about better and more nuanced ways of summarizing research findings.

Neglect of effect magnitude and uncertainty measures

Statistical significance combines two underlying pieces of information: the effect estimate and its uncertainty. However, because statistical significance is used as a filtering mechanism, investigators typically do not emphasize the "uncertainty" component, even though the underlying machinery is of probabilistic nature. As a result, in practice, a statistically significant result tends to be treated as "real, with zero uncertainty." In addition, a nonsignificant result is often interpreted as showing the absence of an effect, as opposed to representing the lack of sufficient evidence to overturn the null hypothesis, despite repeated warnings against such conclusions in statistical textbooks and training. While these two issues are interpretational problems, they occur so often with the null hypothesis significance testing paradigm that they have almost become part of the paradigm itself, making it easy to fall into these conceptual traps.

Some of the above issues can be illustrated by considering NARPS. Given the findings from the 70 independent teams, NARPS performed two types of meta-analyses: one with binarized team reports (logistic regression) and another solely based on statistical values. In the binarized case, the result of each individual team was considered either present (value of 1: the presence of strong evidence is interpreted as an evidence with no uncertainty) or absent (value of 0: the absence of evidence is equated to an evidence of absence). NARPS simply interpreted their meta-analytic findings as indicating substantial variability in study results across different analytical pipelines. A well-known problem with the dichotomization approach is that it treats *p*-values of 0.049 and 0.051, for example, as categorically distinct. On the one hand, the difference between a statistically significant result may not be significantly different from a statistically insignificant one (Fig. 4, Case 1). On the other hand, possibly less appreciated is the fact that the approach neglects differences between the two results that are deemed significant (i.e., in both cases $p < 0.05$) (Fig. 4, Case 2): two results with the same amount of statistical evidence may have nontrivial differences in both effect magnitude and uncertainty. These examples illustrate the extent of information loss due to the sole emphasis on statistical evidence while deemphasizing effect magnitude as routinely practiced in neuroimaging.

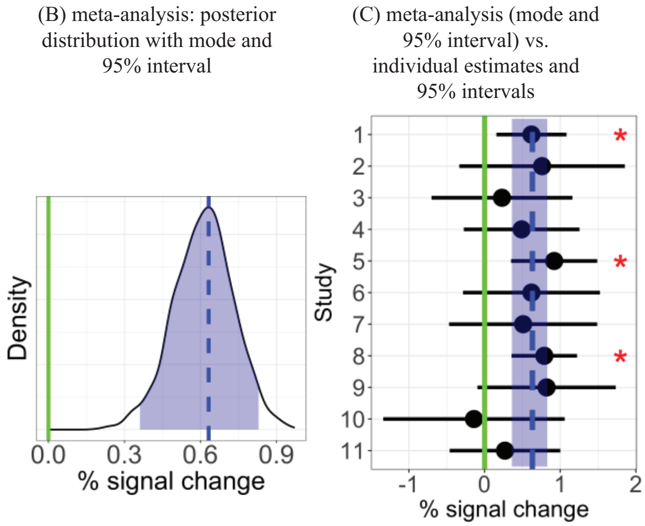
To further appreciate the above issues, consider the hypothetical scenario illustrated in Fig. 5. The example could refer to a series of studies that investigated a specific experimental paradigm in the past (e.g., the contrast of activation in the amygdala between fearful and neutral faces) or to the case considered by NARPS in which different teams analyzed the same dataset. In the scenario, 3 out of 11 results survive the conventional threshold cutoff (Fig. 5A); one may claim poor reproducibility and "sizeable variation" across individual results and question the statistical evidence provided by the suprathreshold studies. This situation only worsens if one imposes an adjustment for multiple testing due to having 11 parallel inferences: with an adjustment applied, none of the studies would survive dichotomization.

Instead of a logistic regression based on binarized assessments, an integrative meta-analysis can be performed by combining the full results of *both* the effect estimate and uncertainty from each study. Just as analytical results vary across different studies, so we should not be surprised to see some extent of variation when the same data are analyzed by different teams unless a high consensus is reached in the field regarding the bolts and nuts of analytical pipelines. On the other hand, to reduce inferential biases and distortion, a proper conclusion regarding the variability would be better achieved through a model with accurate information flow, even if "sizeable variation" is a cause for concern. Specifically, rather than artificially dichotomizing the continuous spectrum of analytical results or reducing their effect estimates and uncertainties to statistical values only (e.g., *t*), we intend to incorporate the full available information into the model.

Let us assume that the effect estimates, $\hat{y}_i$ ($i = 1, 2, ..., 11$), are normally distributed $\hat{y}_i \sim N(\theta_i, \hat{\sigma}^2)$ with mean θ_i and variance $\hat{\sigma}_i^2$. In addition, suppose that the

(A) Individual studies:
effect estimates and uncertainty

Study	$\hat{y}_i$	$\hat{\sigma}_i$	95% interval
1	0.62	0.23	(0.15, 1.09)*
2	0.76	0.54	(-0.36, 1.88)
3	0.23	0.46	(-0.72, 1.18)
4	0.49	0.38	(-0.29, 1.27)
5	092	0.29	(0.34, 1.50)*
6	0.62	0.44	(-0.30, 1.54)
7	0.51	0.51	(-0.49, 1.51)
8	0.79	0.22	(0.35, 1.23)*
9	0.82	0.46	(-0.11, 1.75)
10	-0.14	0.70	(-1.36, 1.08)
11	0.27	0.39	(-0.48, 1.02)


(B) meta-analysis: posterior distribution with mode and 95% interval

(C) meta-analysis (mode and 95% interval) vs. individual estimates and 95% intervals

Fig. 5. Meta-analysis example. (A) Hypothetical results of 11 studies analyzing the same data (or 11 studies of the same task), with results summarized by the estimate of the effect, $\hat{y}_i$ (where *i* is the study index), and its standard error, $\hat{\sigma}_i$. A total of 3 out of 11 effects would be deemed statistically "significant" (red asterisks) according to standard cutoffs. From this perspective, one might say there is inconsistency or "considerable variability" of study results. (B) A different picture emerges if the same studies are combined in a meta-analysis: the overall evidence (area under the curve to the right of zero) points to a positive effect. The posterior distribution of the effect based on Bayesian multilevel modeling provides a richer summary of the results than (A). The shaded blue area indicates the 95% highest density interval (0.36, 0.83) surrounding the mode 0.63 (dashed blue line). (C) The individual results from (A) are presented (dots indicate $\hat{y}_i$, horizontal lines show the uncertainty intervals of one standard error $\hat{\sigma}_i$, and red asterisks mark the individually "significant" studies), along with the meta-analysis distribution information (colors as in B). With the full information present, we can evaluate the study consistency and overall effect more meaningfully.

effects themselves, θ_i, follow $\theta_i \sim N(\mu, \tau^2)$ with mean μ and variance τ^2. The latter distribution specifies a prior and provides some information to the process but only minimally: it assumes that the effects θ_i tend to have a bell-shaped, not uniform, distribution, with some values more likely than others. Under this modeling perspective,[a] we obtain a posterior distribution of μ (Fig. 5B) with the effect at a mode $\hat{\mu} = 0.63$ and a 95% uncertainty interval (0.36, 0.83). When this posterior uncertainty interval is reviewed together with the estimates and uncertainties of the 11 individual studies (Fig. 5C), we now have a convenient way to check and evaluate the consistency of the studies; the fact that the majority of the individual effect mean values fall within (or just outside) the meta-analysis's 95% interval indicates a large degree of consistency, rather than a dichotomized assessment with 3 out of 11 "statistically significant" results.

The last result leads to a very different conclusion than when the meta-analysis was based only on binarized statistics, because the proposed analysis uses both the effect estimate and uncertainty of each individual result. Note that the binarized version is highly sensitive to the definition of "significance" used for the individual studies, as well as to the specific adjustment for multiple testing. Clearly, there is considerable information loss in the processes of binarization and multiple testing adjustment. As an alternative, consider having access only to a summary statistic (e.g., Student's *t*) for each study. A statistic is in essence the ratio of the estimated effect relative to its variability and reduces the two independent pieces of information into one. Whereas using statistic values in meta-analysis is a step in the right direction, it is an insufficient one. Incorporating both the effect estimate and its variability would provide richer information than a statistic value alone. To see this, consider the simple meta-analysis model described above, where the overall effect estimate for *n* studies, given τ, can be stated as[20,21]

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2 + \tau^2} \hat{y}_i}{\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2 + \tau^2}}, \qquad (1)$$

with a standard error $\left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2 + \tau^2} \right)^{-\frac{1}{2}}$. In other words, the full results of the *n* studies are combined through the weighted average among the individual effects $\hat{y}_i$ with their associated variances $\hat{\sigma}_i^2$, together with the cross-study variance τ^2, inversely playing the role of weighting.

The preceding meta-analysis illustrates the value of reporting *both* effect estimate *and* uncertainty values in scientific communication. As FMRI signals do not follow a ratio scale with a true zero, we recommend reporting percent signal change or another index of magnitude, whenever possible. As seen in this section, not providing this information amounts to considerable data over-reduction that leads to many subsequent issues.[22] Reporting effect estimates also helps safeguard against potentially spurious results. Signal changes in FMRI are relatively small and do not surpass 1–2%, except when simple sensory or motor conditions are contrasted with low-level baselines. In contrast, statistical values are dimensionless and do not directly provide information regarding effect magnitude. Indeed, the same statistic value may correspond, for example, to infinitely many possible pairs of mean and standard error (Fig. 4, Case 2). A small *t*-statistic value could represent a small effect with a small standard error or a large effect with a large standard error – two scenarios with very different meanings. In addition, if, for example, a seemingly reasonable statistical value (e.g., *t*-value of 4.3) corresponds to an unphysiological 10% signal change, the conventional "statistic-only" reporting mechanism does not offer an easy avenue to identify and filter out such a spurious result. We note that another common practice of reporting a standardized "effect size" (e.g., Cohen's *d*) shares the same problem of information loss due to (a) the unavailability of the physical scale and (b) the over-reduction from two values (effect estimate and uncertainty) to one. On the other hand, such a standardized metric, if desirable, could be easily derived from the effect estimate and its uncertainty.

The maturation of a field requires some extent of quantification and uncertainty assessment. However, currently, most studies in neuroimaging only report binarized results at the level of qualitative assessment (e.g., positive or negative) with neither specifically quantified effect estimation nor uncertainty. Returning to the NARPS investigation, they performed a second meta-analysis solely based on statistic values. Under this approach without dichotomization, their findings were substantially more consistent with one another across teams, reaching a conclusion that was different from their first meta-analysis based on individual teams' dichotomized reporting. These results are not only encouraging for the field of neuroimaging, but they also highlight the perils of the dichotomous approach. We conjecture that their meta-analysis results would have been further improved if both effect magnitude and uncertainty information had been incorporated in their meta-analyses. On the other hand, their conclusion bias would have been further exacerbated when statistical values were thresholded with "statistically nonsignificant" ones unreported and hidden.

The NARPS investigation, as a prototypical example in the field, highlights the importance of result reporting. For the primary study of interest, the analysis and modeling were set up for massively univariate analyses:

[a] See https://afni.nimh.nih.gov/pub/dist/doc/htmldoc/tutorials/meta/basic_bml.html for the example data and short R code used to perform this example meta-analysis.

inefficient modeling occurred because the information was not shared globally across the brain, and adjustment for multiplicity was necessary. Results were required to be dichotomized in the form of "yes/no" decisions for a few specific regions; model comparison and validation were not part of the common practice. Only statistical evidence was reported in the results, ignoring the informational context of effect magnitudes. The information loss due to these requirements, which mirror many conventional practices, can best be seen in the NARPS report by the contrasting conclusions these steps produce in one meta-analysis with dichotomized value of 0s and 1s compared to another done by looking at unthresholded statistics: the former "resulted in sizeable variation in the results of hypothesis tests," while the latter "analyses of the underlying statistical parametric maps on which the hypothesis tests were based revealed greater consistency than would be expected from those inferences, and significant consensus in activated regions across teams was observed using meta-analysis." That is, the consistency of results was noticeably greater just by loosening one of the sources of information waste (dichotomizing). Similar to the demo example in Fig. 5, it is likely that the finding of more consistency across teams represents reality more closely than the dichotomized version (which had undergone much greater information reduction before assessment). Much of the focus on the NARPS results has been on the "sizeable variation" of the dichotomized results; this has overshadowed the "significant consensus" that was present when the results were shown with less information loss. Due to the common practice of dichotomization and incomplete result reporting, meta-analysis in neuroimaging is largely limited to anatomical locations without regard to effect magnitude and is vulnerable to publication bias. Thus, information loss has a far-reaching impact not only on meta-analysis specifically but also on reproducibility in general.

To conclude this section, let us consider some of the issues discussed in the present and preceding sections. The common statistical practice in population-level analysis faces several challenges:

1. The principle of insufficient reason (9), while reasonable in some statistical settings, disregards distributional information concerning effect magnitude across the brain (Fig. 3).
2. Hard thresholding carries with it a fair amount of arbitrariness and information waste.
3. The use of summary statistics alone to report results instead of a combination of effect estimate and uncertainty has detrimental impacts on study reproducibility (and makes spotting spurious results less straightforward).

Next, we describe how Bayesian multilevel modeling provides a paradigm to address these issues.

Bayesian multilevel modeling

We start with building up the structure of Bayesian multilevel modeling by first considering simple data y_{ij} ($i = 1, 2, ..., n$; $j = 1, 2, ..., k$) from n subjects that are longitudinally measured at k time points with a predictor x_{ij}, using the form $y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$ with intercept α_i, slopes β_i, and residuals ε_{ij}. To appreciate the flexibility of the approach, this formulation is sometimes referred to as a "varying-intercept/varying-slope" model akin to those commonly adopted in a multilevel framework:

$$\begin{aligned}
&y_{ij} \sim N(\mu_{ij}, \sigma_\varepsilon^2);\\
&\mu_{ij} = \alpha_i + \beta_i x_{ij};\\
&\alpha_i \sim N(\alpha, \sigma_\alpha^2), \beta_i \sim N(\beta, \sigma_\beta^2);\\
&\alpha, \beta \sim N(0, 1);\\
&\sigma_\alpha, \sigma_\beta, \sigma_\varepsilon \sim \text{Half-Cauchy}(0, 1).
\end{aligned} \tag{2}$$

What makes the model "multilevel" is that it involves the hierarchical levels of subjects and time points. The notation α_i indicates that each subject i has a unique intercept; likewise, β_i shows that each subject i is given a unique slope. The first line specifies the *likelihood* or the distributional assumption for the data y_{ij}. The expression for μ_i specifies a linear relationship with a single predictor x_{ij} (adding more predictors is straightforward). The third line shows the *priors*: the varying intercepts follow a Gaussian distribution with a grand intercept α plus standard deviation σ_α; likewise, the varying slopes follow a Gaussian distribution with a grand slope β plus standard deviation σ_β. Finally, the last two lines specify the *hyperpriors* (parameters of the prior distributions), which can be conveniently weakly informative distributions for the means and standard deviations of the priors.

The above Bayesian multilevel modeling framework can be applied quite generally to any hierarchical structure. For example, meta-analysis is typically formulated under the conventional framework, as shown in formula (1), through random-effects modeling. However, it can also be conceptualized as a Bayesian multilevel model as exemplified in Fig. 5. Even though the two approaches would often reach similar conclusions except for some degenerative cases,[b] the posterior distribution from Bayesian modeling provides richer information than a point estimate combined with a standard error. As illustrated in Fig. 5, we do not assume a uniform prior by adopting the principle of insufficient reason nor do we adjust for multiple testing for individual studies as in the massively univariate approach. Rather, we regularize or

[b] For example, a zero variance estimate ($\tau^2 = 0$) may arise under the conventional framework, especially when the number of studies is small. Such an implausible boundary estimate would not occur under the Bayesian formulation (23).

apply partial pooling on the studies through weighting as shown in formulation (1).

The Bayesian formulation (2) allows the modeler to flexibly estimate intercepts and slopes as a function of the hierarchical level of interest. Due to the impact of partial pooling across hierarchical levels, the Bayesian model tends to generate estimates that are more conservative and closer to the average effect within a given hierarchy than if each specific effect was estimated individually. Because of this conservative nature, the multilevel model aims to control for errors of incorrect magnitude (type M) and sign (type S). Furthermore, adjustment for multiplicity is not needed (24), especially since all the inferences are drawn from a single, overall posterior distribution of an integrative model.

In the past years, we have investigated how the Bayesian framework can be effectively employed to analyze FMRI data at the region level (8,25,27), as well as for matrix-based analysis including time series correlations among regions (28). The approach has also been applied effectively to other scenarios in neuroimaging.[26,29–31] Although at present the framework is computationally prohibitive at the whole-brain voxel level, we have also employed the technique at the voxel level within brain sectors, such as the insula. First, we present an example that illustrates a recent application (25) at the region level. The outcome in Fig. 6A shows the probabilities of observing the effect of interest in a range (for example, a positive effect, or a negative effect). For each region, the full posterior distribution conveys both the effect magnitude and its uncertainty, and the latter can be captured by the area under the curve to the right/left of zero. This posterior can be reported in full without dichotomization, as shown here. For example, the posterior probability that the effect was greater than zero in the left superior frontal gyrus (LSFG) was 0.92, which may be noteworthy in the research context in question. In particular, model fits can be qualitatively assessed by plotting predicted values against the raw data through posterior predictive checks (Fig. 6B) and quantitatively compared to alternative models using information criteria through leave-one-out cross-validation. By comparison, the model fit using the massively univariate approach was considerably poorer (Fig. 6B).

The Bayesian multilevel approach can also be applied to voxel-level data within spatially delimited sectors. For instance, in a recent experiment, two separate groups of participants received mild electrical shocks (31). In the *controllable* group, participants could control the termination of shocks by pressing a button; in the *uncontrollable* group, button pressing had no bearing on shock duration. The two groups were yoked so that, for a given participant, the exact timing of shock events in the controlled condition was replicated for a paired participant in the uncontrolled condition. In the standard FMRI approach, at the voxel level, the effects (commonly denoted as β coefficients) of each participant were estimated based on a time series regression model; one would proceed with voxelwise inferences (say, a *t*-test comparing the two groups) followed by a threshold adjustment based on the spatial extent to control for multiple testing.

In contrast, the Bayesian multilevel approach specifies a single model, which combines all data according to natural hierarchical levels of the data. In this particular study, one natural hierarchy was that of participant pairs given the yoking of the experimental design. In addition, we focused on voxels within the insula, a cortical sector important for threat-related processing. However, the insula is a large and heterogeneous territory, with notable subdivisions that previously had been

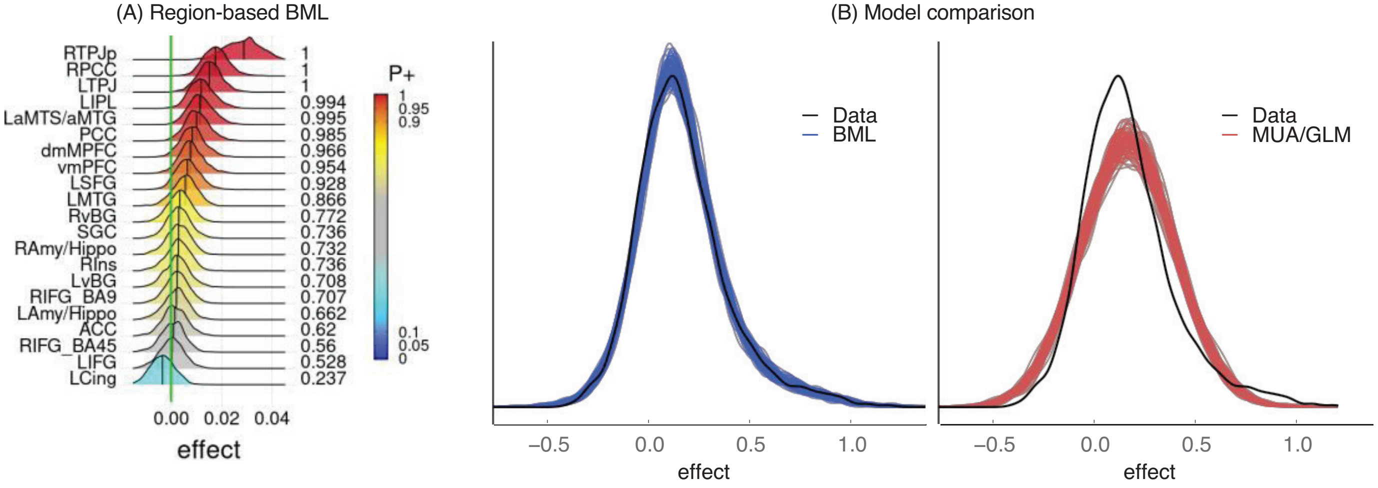


Fig. 6. Bayesian multilevel (BML) modeling at the region level. (A) Population-level analysis was performed with an FMRI study of 124 subjects.[25] Each curve shows the posterior distribution (probability density). Colors represent values of P^+: the posterior probability that the effect is positive. The analysis revealed that over one-third of the regions exhibited considerable statistical evidence for a positive effect. In contrast, with massively univariate analysis, only two regions survived multiple testing adjustment.[26] (B) The BML performance was assessed and compared to the conventional approach.[25] Posterior predictive checks visually compare model predictions against raw data. The BML model generated a better fit to the data compared to the general linear model (GLM) used in the massively univariate analysis (MUA).

described functionally and anatomically. Accordingly, we subdivided the insula in each hemisphere into 10 subregions, each with approximately 100 voxels. Thus, the subregions comprised another hierarchy. At the most basic level of the hierarchical structure, the unit was the voxel itself.

The following model was employed for the voxel-level data,

$$\Delta_{prv} \sim N(\mu_{prv}, \sigma_\varepsilon^2),$$

$$\mu_{prv} = \alpha + \beta_p + \gamma_r + \theta_v,$$

where the difference Δ_{prv} in BOLD responses to shock between a participant pair *p* in a voxel *v* belonging to region *r* was estimated at the subject level through time-series regression analysis and assumed to originate from a Gaussian distribution centered on μ_{prv} with variance σ_ε^2. The second line specifies the response difference at the population level as a linear combination of an overall effect α, a contribution β_p from participant pair *p*, a contribution γ_r from region *r*, and a contribution θ_v from voxel *v*. Importantly, the participant pairs, regions, and voxels are assumed to come from their respective (hypothetical) populations characterized by priors as in model (2) (further specifications omitted here for brevity) and play a role equivalent to "random effects" in conventional linear mixed-effects models. Finally, for simplicity here, we omitted several covariates that were included in the original analysis, including those related to individual differences in trait and state anxiety. Those covariates can be captured by slope parameters as in model (2), where it is possible to include varying slopes (thus slopes can vary across regions, for example). This Bayesian machinery allowed us to estimate the contributions of participant pairs, regions, and voxels based on the data, the likelihood, and the prior distributions. In this study, our goal was to understand voxelwise effects (Fig. 7).

To recapitulate, we note that the Bayesian approach can be adopted to achieve seven important goals.

1. *Handling multiplicity.* The Bayesian framework offers a potential avenue to addressing the problem of multiple testing that is so central to neuroimaging statistics. Because a *single* model is employed with information shared and regularized through partial pooling, all inferences are drawn from a single overall posterior distribution. Thus, information is more efficiently shared across multiple levels; no multiple testing adjustment is necessary (24), avoiding excessive penalty due to information waste. In other words, instead of resorting to a post hoc adjustment for multiple testing under a modeling framework with an unrealistic assumption (e.g., uniform distribution), the Bayesian approach directly incorporates the interrelationships of the hierarchical structure as part of the modeling processing. We note that some statisticians have suggested other forms of adjustment based on decision theory (2,32,33).

2. *No penalty against small regions.* Under massively univariate analysis, the spatial extent is traded off against voxel-level statistical evidence in the process of adjusting for multiple testing. Thus, small regions are inherently placed in a disadvantageous position even if they have similar effect strength as larger ones. In contrast, under the Bayesian framework, each spatial unit is a priori assumed to be exchangeable from any other units. In other words, all units are a priori treated on an equal footing under one common prior distribution and are a posterior assessed on their own effect strength. As a result, small regions are not disadvantaged because of their anatomical size (8).

3. *Insensitivity to data space.* Under the conventional framework of massively univariate analysis, the statistical evidence is sensitive to the amount of data: when the investigator confines their focus to a particular region instead of the whole brain, the statistical evidence would suddenly become "stronger" even though the data remain

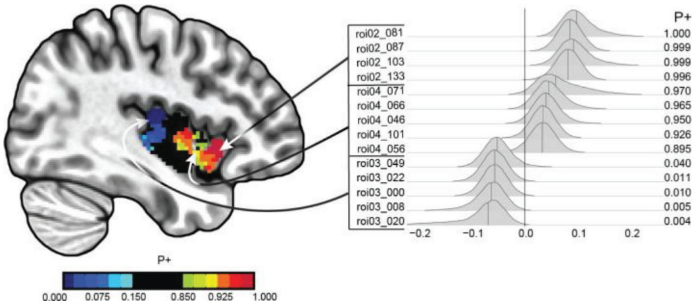


Fig. 7. Bayesian multilevel voxelwise results. The right part of the figure illustrates posterior distributions of voxels from three subregions of the insula (voxels selected to illustrate some of the range of statistical evidence). Colors represent values of *P+*: the posterior probability that one condition (uncontrollable group) is greater than the other (controllable group). Values closer to 1 indicate stronger evidence that uncontrollable is greater than controllable, while values closer to 0 indicate the opposite (values computed based on the posterior distributions of the difference of the two conditions correspond to the tail areas of the posteriors). The computational time was about 2 weeks for this dataset with 126 subjects and approximately 1,000 voxels on a Linux server using four Markov chains.

the same (e.g., "smaller volume correction"). In contrast, under the single integrative framework, the information is shared in a "melting pot" and calibrated. In other words, partial pooling plays a self-adaptive role of regularization, similar to the situation with the conventional methods such as ridge regression and LASSO (least absolute shrinkage and selection operator). Thus, the impact on the same spatial unit is relatively negligible even when the total amount of data changes (e.g., increasing or decreasing the number of spatial units); that is, a region would be assessed by its own "merits" of effect magnitude but not its anatomical size (8).

4. *Model quality control.* For various reasons, model performance and comparisons are rarely cross-examined in neuroimaging. However, the modeling process should not be simply executed as an automatic pipeline without quality control. In fact, model accuracy and adequacy can be assessed through posterior predictive checks and cross-validation under the Bayesian framework. For example, a posterior predictive check allows one to examine the model adequacy or discrepancies through visually comparing the predictive distribution to the observed data. Cross-validation is another important technique under the Bayesian framework to gauge how closely a model, relative to potential candidates, predicts future data from the same data generating processes that produced the current data at hand. In general, the Bayesian approach welcomes an integrated view of the modeling workflow with an iterative process of model development and refinement (34).

5. *Enhanced interpretability.* The Bayesian approach enhances the interpretability of analytical results. For instance, the posterior probability indicates the strength of the evidence associated with each effect estimate, conditioned on the data, model, and priors. In the conventional null hypothesis framework, uncertainty is expressed in terms of standard error or confidence interval. Unfortunately, while mathematically precise, this information is very difficult to interpret in practice and easily misunderstood (35). Notably, a confidence interval is "flat" in the sense that it does not carry distributional information; parameter values in the middle of a confidence interval are not necessarily more or less likely than those close to the end points of the interval, for example (e.g., Fig. 5A, C). In contrast, the posterior distribution provides quantitative information about the probability of ranges of values, such as the parameter being positive, negative, or within a particular interval. Unlike the conventional notion of "confidence interval," parameter values surrounding the peak of the posterior distribution are more likely than those at the extremes (Figs. 5B, 6A, 7).

6. *Error controllability.* Instead of the false positive and false negative errors associated with the conventional null hypothesis significance testing framework, Bayesian multilevel modeling can be used to control two different errors: *type M* (over- or under-estimation of effect magnitude) and *type S* (incorrect sign) (36). For example, effect estimates under the massively univariate modeling framework tend to be exaggerated (Jensen's inequality), leading to type M error. In contrast, shrinking the effect estimates under the Bayesian multilevel framework provides an effective way to regularize and counteract the impact of exaggeration (14).

7. *Extended modeling capabilities.* The Bayesian framework is advantageous and flexible in handling complex data structures that can be challenging for the conventional framework. Consistent with the central limit theorem, many types of data, including the effect estimates in neuroimaging, tend to have a density of roughly Gaussian characteristics with a bell-shaped distribution exhibiting a single peak and near symmetry. Based on the maximum entropy principle, the most conservative distribution is the Gaussian if the data have a finite variance (37). Thus, for the same reasons that subjects in neuroimaging are routinely treated as random samples from a hypothetical pool of a Gaussian distribution, we can effectively model the effect distribution across space as a Gaussian, rather than adopting the stance of "full ignorance." However, exceptions do occur when the data do not follow a bell-shaped distribution due to outliers or skewness. A conventional approach is to set hard bounds (e.g., the rule of three standard deviations), constraining the data to a predetermined interval in order to exclude outliers. Such a brute-force approach is arbitrary and unprincipled to some extent. In contrast, outliers or skewed data can be accommodated in a principled manner with the utilization of non-Gaussian distributions (e.g., Student *t*-distribution with an adaptive number of degrees of freedom, Lambert W transforms) for data variability. Another benefit of Bayesian modeling is the convenience of incorporating the uncertainty information (e.g., measurement errors) for both response and explanatory variables that may improve modeling accuracy and accommodate data asymmetry due to outliers or data skewness.

Theoretically, the Bayesian multilevel framework can incorporate any number (large or small) of spatial units, voxels, or regions into one unified model. Numerical considerations aside, such a Bayesian model is essentially the same as the traditional linear mixed-effect formulation both in conceptual viewpoint and in symbolic model expression. In addition, the crucial aspect of the hierarchical framework lies in the assumption of, for example, a Gaussian distribution for the variability across spatial units. It is this distribution assumption that plays the role of information sharing across space through global calibration or shrinkage. If the prior is relaxed from a Gaussian distribution to a trivial case of uniform distribution, then no information is shared across the spatial units per the principle of indifference that assigns epistemic probabilities (9). In other words, in the absence of available evidence, one could adopt a uniform distribution across space; thus, the hierarchical model would simply reduce to a special case, namely, the conventional massively univariate model. However, as discussed here, it is generally more reasonable as well as more informationally efficient to adopt a hierarchical model with the assumption of an approximately bell-shaped, rather than uniform, distribution for cross-spatial variability, as evidenced in the empirical data of Fig. 3B.

The adoption of Bayesian multilevel modeling here is intended to specifically address information waste through three issues at the population level: mischaracterization of data hierarchies, dichotomization, and data over-reduction. There has been a rich literature of Bayesian applications in neuroimaging that focus on various aspects of modeling. The wide range of topics include the following: using Bayesian modeling as an alternative at the subject level (e.g., temporal structure (38), spatiotemporal modeling (39), complex-valued FMRI (40), hemodynamic response estimation and connectivity (41), empirical Bayes for spatiotemporal modeling (42–48)), adopting empirical Bayes to resolve the unstable cross-subject variability (e.g., when the number of subjects is 40 or less) by sharing data variability across neighboring voxels (49), leveraging between anatomical data with a high resolution and those with a low resolution (50), handling measurement errors (51). Some of these methods have adopted a similar concept of partial pooling for time series regression at the subject level (42,44,45,48) or for locally regularizing cross-subject variability at the population level (49). However, it is beyond the scope and space in this commentary to provide an exhaustive and detailed coverage for these topics that are tangential to our focus of addressing the issue of information loss at the population level.

Neuroimaging without *p*-value thresholds?

Let us consider the issue of probability thresholding, regardless of the modeling framework, in further detail. Dichotomization is essential to statistically based decision-making. As noted above, it provides a way to filter a lot of information and to present results in a highly digestible form: binary ON/OFF output. For example, based on the available data, should a certain vaccine be administered to prevent Covid-19? In such cases, a binary decision must be adopted, and decision theory, which incorporates the costs of both false positives and false negatives, can be used. Here, we entertain a seemingly radical proposal: What would be lost in neuroimaging if hard thresholds were abandoned? It could be argued that this would lead to an explosion of unsubstantiated findings that would flood the literature. We believe this is unlikely to occur. Scientists are interested in finding the probability of seeing the effect conditioned on the data at hand ("what happened"), rather than the *p*-value ("what might have happened" or the probability of seeing the data or more extreme scenarios conditioned on the null effect). The absence of a hard threshold does not entail that "anything goes"; rather, it encourages substituting a mechanical rule by the careful justification of the noteworthiness of the findings in a larger context. Consider the controllability study discussed above. In additional analyses at the level of brain regions, we found very strong evidence (*P*+ = 0.99) for a controllability effect in the bed nucleus of the stria terminalis, a structure that plays an important role in threat processing. This region and the central nucleus of the amygdala are frequently conceptualized as part of a functional system called the "extended amygdala." Accordingly, we found it important to emphasize that there was also some evidence (*P*+ = 0.90) for a controllability effect in the left central amygdala. Although the central amygdala did not meet typical statistical cutoffs, we believe that the finding is noteworthy in the larger context of threat-related processing. This is particularly the case because reporting the central amygdala effect can be informative when integrating it with other studies to perform a meta-analysis, as discussed earlier. Note that by providing the information about the central amygdala, readers are free to interpret the findings in whatever way they prefer; they may agree with our interpretation (that there is some evidence for an effect in this region) or consider the evidence "just too weak." This is not a problem in our view; rather, it is a feature of the approach we advocate for.

A more flexible approach both in terms of statistical modeling and in terms of result reporting is potentially beneficial. At the heart of the scientific enterprise is rigor. In experimental research, typically this translates into testing patterns in data in terms of null hypotheses and a *p*-value threshold of 0.05. On the surface, the precise cutoff provides an objective standard that reviewers and journal editors can abide by. On the other hand, the use of a strict threshold comes with its own consequences. In most research areas, including neuroimaging, data are notoriously variable and not readily accommodated by simple models (37). In this context, is it really essential to treat a cluster size of, say, 54 voxels as qualitatively different from one with 50 voxels? As models by definition have limitations, we believe that dichotomization, as illustrated by the example in Fig. 5, is unproductive.

In light of these considerations, we propose a more "holistic" approach that integrates both *quantitative* and *qualitative* dimensions. A recent investigation through Bayesian multilevel modeling indicates that full result reporting including visualization can effectively replace dichotomous thinking (52). For results based on the conventional framework, we suggest a general *highlight but not hide* approach. Instead of applying a threshold that excludes results that do not cross it, one can show all (or most) results while highlighting or differentiating different levels of statistical evidence (53) (Fig. 1F). Similarly, tables can include regions with a broad spectrum of statistical evidence, together with both their effect magnitudes and uncertainties. Overall, probability values, including the conventional *p*-value based on null-hypothesis testing, play a role as a piece of information, rather than serving a gate-keeping function. In addition, we encourage a mindset of "accepting uncertainty and embracing variation" (54) in the results of any particular study.

Modeling trial-by-trial variability

In this section, we further illustrate the potential of using the Bayesian multilevel approach to build integrative analysis frameworks. In FMRI experiments, the interest is usually on various comparisons at the condition level. As condition-level effects exhibit considerable variability, researchers rely on multiple trial repetitions of a given condition to estimate the response via a process that essentially amounts to averaging. In this manner, trial-by-trial variability is often treated as noise under the assumption that a "true" response exists, and deviations from it constitute random variability originating from the measurement itself or from neuronal/hemodynamic sources.

However, neglecting trial-by-trial variability means that trial-level effects are considered as "fixed" in the fixed-versus random-effects terminology, as opposed to participants, who are treated as random and sampled from a hypothetical population. Technically, this means that researchers cannot generalize beyond the specific stimuli employed in the experiment (say, the 20 faces used from a given dataset), as recognized several decades ago (55,56). By modeling trials as varying instantiations of an idealized condition, a study can generalize the results to trials beyond the confine of those employed in the experiment (57,58). Consider a segment of a simple experiment presenting five faces. In the standard approach, the time series is modeled with a single regressor that takes into account all face instances (Fig. 8A, B). The fit does a reasonable job at capturing the mean response; however, it is clearly poor in explaining the fluctuations at the trial level (Fig. 8C). Whereas traditional models in neuroimaging ignore this variability across trials, we propose to explicitly account for it in the underlying statistical model (57,58).

The Bayesian multilevel framework can directly be used to account for trial-level effects. Specifically, at the subject level, we construct regressors for individual trials as in Fig. 8D. In a recent study, we explored a series of population-level models of trial-by-trial variability for FMRI data (58) and indeed observed considerable cross-trial variation and notable inferential differences when trials were explicitly modeled. For example, as the experiment

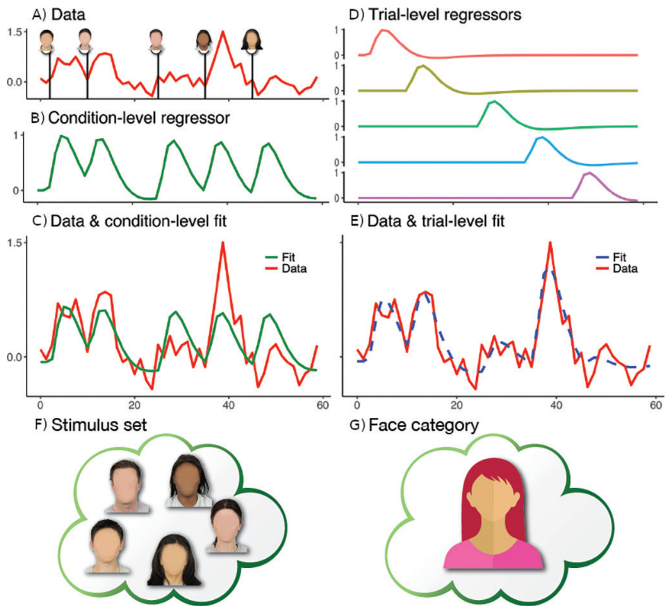


Fig. 8. Time series modeling and trial-based analysis. Consider an experiment with five face stimuli. (A) Hypothetical times series. (B) The conventional modeling approach assumes that all stimuli produce the same response, so one regressor is employed. (C) Condition-level effect (e.g., in percent signal change) is estimated through the regressor fit (green). (D,E) Trial-based modeling employs a separate regressor per stimulus, improving the fit (dashed blue). (F,G) Technically, the condition-level modeling allows inferences to be made at the level of the specific stimulus set utilized, whereas the trial-based approach allows generalization to a face category.

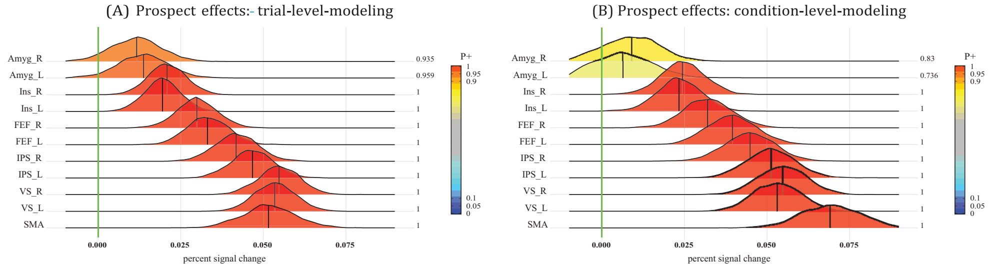


Fig. 9. Trial-level versus condition-level modeling. Posterior distributions for the effect of reward (vs. control) cues for each region of interest. Although the two approaches provided comparable results, trial-level modeling (A) showed stronger evidence for the left and right amygdala than the condition-level counterpart (B).

included a task involving negative or neutral faces, we were interested in amygdala responses, but our interest extended to a trial phase only containing cues indicating whether the trial was rewarded or not (in reward trials, participants received extra cash for correct and timely responses). Fig. 9 shows that trial-level modeling provided considerably stronger evidence for an effect of reward in the amygdala compared to the conventional condition-level modeling.

Trial-level modeling also improves the estimation of test–retest reliability (i.e., the degree of agreement or consistency of subject-level differences between two or more repeated measurements). Recent reports have suggested that the test–retest reliability for psychometric (59) and neuroimaging (60) data is rather low when evaluated via the conventional intraclass correlation coefficient. The low reliability of effects with robust population-level effects (e.g., Stroop and Flanker tasks) was particularly worrisome in the context of individual differences research. In a recent study, we developed a multilevel modeling framework that takes into account the data hierarchies down to the trial level, providing a test–retest reliability formulation that is disentangled from trial-level variability (27). As a result, the trial-level modeling approach revealed the attenuation when the conventional intraclass correlation coefficient is adopted and improved the accuracy of reliability estimation in assessing individual differences.

Two complex issues about trial-level modeling are experimental design and trial sample size. When the effect at each trial is separately characterized in the model at the subject level, high correlations or multicollinearity may arise among the regressors. To avoid such potential issues, trial sequence and timing can be randomized in the experimental design. As shown in a few recent studies (27,58,61), even fast event-related experiments with a short inter-trial interval can be carefully designed so that trial-level effects can be captured. Nevertheless, detailed attention is still needed in processing and quality control, because unstable effect estimates, outliers, and skewed distributions may still occur due to high collinearity among neighboring trials or head motion. Our recent investigations (27,58,61) provide some solutions to handle such difficult situations. Furthermore, even though the trial sample size is largely chosen as a convenient or conventional number with which the subject would be able to endure during the scanning session, one study (61) indicates that it has nearly the same impact as subject sample size on statistical efficiency.

DISCUSSION

Neuroimaging research is challenging, not least because data analysis includes several interdependent steps of processing and modeling. Data from tens of thousands of spatial units are acquired as a function of time for one or multiple subject groups and for several experimental conditions with trials repeated many times per condition, typically across multiple data acquisition runs. Given the challenges any one research team would face to analyze this type of data, developers have designed software packages that enormously lower the barrier to entry to investigators. Indeed, statistical development for FMRI analysis has proceeded vigorously since the early 1990s. Among the greatest challenges has been the issue of multiple testing, with the dream of "whole-brain noninvasive" imaging coming at a severe cost inferentially. Since the beginning, experimenters have been admonished that without "strict-enough" procedures, the "false positive rate" would be prohibitively high. Accordingly, considerable research has been devoted to improving inferential rigor.

Conditionality of statistical information

Contrary to common practice, the strength and accuracy of statistical evidence are not as informative as usually perceived. According to the central limit theorem, given a large-enough sample size (e.g., big data initiatives), statistical evidence may reach as strong as any designated level unless the associated effect is absolutely zero. For example, suppose that the contrast between positive and negative valences in a brain region follows a Gaussian distribution $N(\mu, \sigma^2)$ with $\mu = 0.2$ and $\sigma = 0.5$ (in units of percent signal change). With a sample size of 40 or 1,000 subjects, one could condense and reduce the data information, as typically done in the literature, to a single number, such as a t-value of 2.7 ($p \approx 0.01$) or 10.0 ($p \approx 1.0 \times 10^{-22}$), respectively. It is difficult to properly digest the information from these two statistical values and to assess as to how much more information is contained in the latter than the former. In contrast, much more revealing information could be attained if their posterior distributions or corresponding 95% uncertainty intervals of the BOLD effect were provided (e.g., (0.04, 0.36) and (0.17, 0.23)).

Statistical interpretation should be properly framed and contextualized. In the aforementioned example, an uncertainty interval only characterizes the population average μ, which is largely a theoretical or abstract construct; thus, one should not lose sight when gauging a particular subject's effect, which could vary in a much larger range (cf., $N(0.2, 0.5^2)$). In addition, any statistical inference is conditional on the adopted modeling framework, the underlying assumptions, and the representativeness of the sampled data. For instance, all statistical models are to some extent an idealization or approximation of the reality; pragmatically, analytical decisions require scientific evaluation, including specific aspects of data processing (amount of spatial smoothing, choice of data included or excluded, model validation, etc.) and uncertainty assignment through a probability distribution. Therefore, estimation accuracy, statistical evidence, and probabilistic reasoning may change if, for example, confounding variables, interaction effects, and nonlinearity are incorporated, when a different distribution is assumed, or if some extent of regularization is applied.

Applicability of Bayesian multilevel framework

Bayesian modeling in general has several advantages that could substantially benefit the neuroimaging community, but some barriers at present hinder its wide adoption. We list here a few directly relevant, but not exhaustive, advantages:

1. Natural, intuitive, and straightforward interpretation based on probabilistic inference,
2. The convenience of incorporating prior information (e.g., distribution across spatial units),
3. The capability of explicitly capturing the data generative mechanism,
4. Efficient handling of multiplicity through partial pooling,
5. No penalty against regions because of their small anatomical size,
6. Strong power of numerical solutions through Monte Carlo simulations,
7. Full result reporting that contains both effect magnitude and uncertainty,
8. Built-in model comparisons and quality check.

In contrast, at least three negative aspects prevent the Bayesian framework from its wide application: two of them are educational and the third practical. First, in most classrooms, the teaching content of Bayesian modeling is sparse and generally quite outdated. As a result, most investigators are not well versed to quickly adopt the concept and structure of the modeling framework. Second, historically there is a negative impression of "subjectivity" associated with the notion of priors. Third, computational hurdles have slowed the uptake of Bayesian applications until the numerical breakthrough of Markov chain Monte Carlo simulations.

Despite these hurdles, we are optimistic on the expanding potential of Bayesian modeling, as each barrier has been decreasing over time. First, there is more awareness and educational momentum about these methods, increasing their popularity, application, and development. For instance, a Bayesian multilevel model usually has a counterpart in the conventional linear mixed-effects formulation and contains the latter as a special case. Thus, the realization of this fact may help resolve some of the difficulties for those who are accustomed to the conventional paradigm. Second, while Bayesian modeling does need to choose priors, one must remember that the conventional massively univariate approach also makes an implicit prior assumption (a uniform distribution across space), leading to information waste. One must also remember that the goal of modeling is to closely characterize the data hierarchies: Bayesian modeling is more powerful in the sense that its performance and the adoption of priors can be rigorously evaluated via tools such as posterior predictive check (Fig. 6) and cross-validation. Finally, in terms of current practical limitations, we are optimistic that the rapid pace of development will increase computational efficiency and speed.

Hierarchical modeling provides a suitable platform for closely characterizing the data generative mechanism, and Bayesian multilevel modeling offers an important numerical machinery in making statistical inferences. Due to the complex and intertwining data structures in neuroimaging that involves many levels such as time series, trials, conditions, subjects, groups, and brain regions, it is pivotal to adopt a holistic perspective that reflects as closely as possible the sources of data variability. It is also this hierarchical consideration that motivates us to propose the incorporation of spatial units as part of the modeling process as opposed to a post hoc compensation for multiple testing under the conventional massively univariate framework. Some of the hierarchical schemata are conceptually equivalent to various regularization methods (e.g., ridge, LASSO); they can be also formulated under the conventional mixed-effects paradigm with, for example, the spatial units playing the role of "random effects." Although computationally affordable, a linear mixed-effects model could only allow one to make inferences at the population level but not for spatial units as individual random effects. In contrast, Monte Carlo simulations, despite the relatively high numerical cost, enable the Bayesian framework to accommodate a wide range of modeling capability and inferential power. We offer three programs of Bayesian multilevel modeling in neuroimaging as part of the AFNI suite for public use: RBA (25) and MBA (28) for region- and matrix-based analyses and TRR (27) for test–retest reliability estimation.

Analytical level: voxel versus region

The choice of analysis level – voxel or region – deserves some elaboration and discourse. Both are commonly used across the neuroimaging field, and the choice involves some trade-offs in processing and modeling at the population level; it can even affect final interpretation (though, some differences are not as large as they might appear).

Modeling at the voxel level has the benefits of relatively high spatial resolution and independence of a choice of region definition. However, as the inferential focus is usually on the cluster – not voxel – level, one thorny issue is the lack of spatial specificity that plagues the massively univariate framework. A post hoc solution (62) has been proposed to address and improve the issue of lacking spatial specificity; this provides an interesting approach, which could also be integrated with other issues raised here (e.g., lack of region-level uncertainty, information loss due to the absence of global calibration). Nevertheless, the real spatial resolution of signals is not actually the acquired voxel size (nor any upsampled final size). There is inherent smoothness in the acquired data and more is added by blurring during preprocessing; this blurring decreases spatial resolution by spreading information across anatomical structures, which makes final interpretation more difficult and less spatially specific. The initial spatial specificity is further lessened by the post hoc adjustment for multiplicity when modeling at the voxel level through massively univariate analysis, which commonly relies on spatial clustering. In the end, individual voxels are not interpretable, and only surviving clusters are. Additionally, voxelwise analyses are not independent of brain regions: the location of voxels within anatomical structures still matters in multiple ways. In practice, most adjustments for multiplicity penalize small regions. Importantly, even voxelwise studies aim to produce and make inferences at the region level. The typical research hypothesis is framed in terms of regions of interest, such as "Based on previous literature, we hypothesize that regions X, Y and Z will show…"As a recent example, NARPS asked teams to report yes/no findings about certain regions of interest, which were then combined for meta-analyses (although a specific atlas was not specified, researchers had to choose their own definitions).

Moreover, voxelwise analyses have been susceptible to problematic or inconsistent result reporting. While these issues are not strictly inherent to the approach, their prevalence makes it difficult to disentangle. Often, there is overfitting at the voxel level, leading simultaneously to exaggerated estimates and poor prediction accuracy. Furthermore, with dichotomization through clustering, interpretational difficulties can arise. Typically, each cluster is not localized within a single brain region – should all the regions be reported, or just those with large overlap (how to define "large"?)? Many researchers just report the location of a single voxel with "peak" statistical values, in order to summarize the results for the entire cluster. This is statistically inconsistent with the thresholding procedures and fundamentally a large source of information waste. The issues of inefficient modeling, artificial dichotomization, and spatial ambiguity further propagate to problems in downstream processing steps. For example, with the goal of categorically determining the spatial intersection of two or more dichotomized clusters, the conventional conjunction analysis is also vulnerable to the aforementioned issues.

Region-based studies have some drawbacks as well. For example, there is a semi-arbitrariness of parcellation selection, which will affect results. Furthermore, not all structures of interest exist in available parcellations (though the number available will surely only increase over time). With multiple regions analyzed through conventional approaches, the penalty for multiplicity is often quite high, to the point of being unacceptable and/or unrealistic. In some cases, subregion differentiation can be important (e.g., localizing focal lesions), but these tend to be clinical cases and do not often occur in population-level studies.

However, there are several useful properties of region-based analyses. First, there is a meaningful specificity to results: in most brain studies, including NARPS, research hypotheses are based on regions, and in this case, the list of relevant regions is clearly defined from the start and consistent with the study design. Blurring does not have to be included in the preprocessing, so signals are not spread widely across regions. Having a smaller number of spatial units means that analyses take less computational time and cost and also that it is easier to avoid dichotomization. Practically, modeling at the region level opens the door to capture the hierarchical structures and to improve model efficiency. The artificial black-or-white classification at the cluster level in the conventional conjunction analysis would be replaced by quantitative characterization through direct effect comparisons at the region level. Finally, all regions – big or small – are treated equally.

All things considered, region-based and voxelwise analyses at the population level share several commonalities. Each relies heavily on alignment, even if in different ways. By the time modeling is complete, the spatial resolution of the two approaches is likely not very different, depending on the parcellation. In fact, one can note that recent parcellations have created more and more regions in a standard template brain (greater than several hundred regions, in some cases) so that their spatial resolution becomes finer. Indeed, at this level, voxels are then just the limiting case of finer parcellations, particularly once blurring has been accounted for. In both cases, what remains most important is to utilize a reasonable modeling framework and statistical practices with either methodology, reducing inconsistency and information waste to the greatest extent possible.

Maintaining enough information in result reporting

The conventional massively univariate analysis as an exploratory tool can benefit from our discussion here regarding the principles and rationales underlying the Bayesian framework. Because of the computational burden, currently, the Bayesian multilevel model can only afford to handle up to a few thousand spatial units (e.g., regions or voxels); thus, exploratory analysis at the whole-brain voxel level is presently beyond the reach of Bayesian modeling. However, future methodological developments and computational breakthroughs will surely continue to reduce, if not eliminate, this computational barrier (63,64). Nevertheless, we believe that the hierarchical perspective helps reveal the information loss associated with two aspects of the conventional modeling approach: the implicit assumption of uniform distribution and the artificial dichotomization required in handling multiplicity. In fact, these two aspects are two sides of the same coin: the conventional modeling methodology focuses only on local relatedness among neighboring spatial units but ignores the global information shared across the whole brain. Consequently, the various approaches of adjustment for multiple testing adopted in the field may lead to excessive penalties and overconservative inferences. For these considerations, when voxelwise analysis is performed under the conventional massively univariate framework, we believe that the Bayesian multilevel framework lends an important perspective: a threshold or a set of spatial blobs purely based on statistical evidence is only suggestive and should not be treated as being etched in stone. To avoid further information waste, any statistical evidence should be viewed – regardless of the adopted framework – as intrinsically embedded with some underlying and implicit assumptions; it should be considered as a continuum both in result reporting as well as during the research reviewing process.

Here, we have addressed a few issues within conventional neuroimaging analysis pipelines: in the process of breaking down raw data and turning it into understandable results, we do not focus on boiling everything down to a small number of ON locations (in a sea of OFF background) at a given statistical significance level. We have shown the many ways that this can be considered an "overdigestible" result: a lot of useful information has been sacrificed (results at subthreshold locations that might still be informative, and separate effect estimates with uncertainty measures) for not much gain. Additionally, we have demonstrated that the conventional modeling approach is inefficient and wastes data, even before getting to questions of dichotomization: the implicit assumption of uniform distribution is far from approximating any realistic brain effect, and *p*-values only provide limited information about how unlikely the current data or more extreme observations would be if a null effect *were* true, rather than the probability of research hypothesis being true *given* the present data.

Instead, we have proposed a small but important improvement to standard neuroimaging pipelines with an approach that aims to make more efficient use of the initial data, and that also has positive side effects for scientific inquiries. A schematic of this approach is shown in Fig. 10, in direct comparison with the traditional approach in terms of information loss and digestibility. First, the Bayesian multilevel modeling approach replaces the massively univariate analysis and the principle of insufficient reason with a single integrative model; the approach also removes any later need for multiple testing adjustment. One benefit of this approach is now obtaining an overall posterior distribution for all model parameters, which provides a great deal of useful information about the estimate uncertainty as well as the overall model fit. This procedure also employs partial pooling across spatial units, so that the effect estimates are regularized to avoid potential overfitting.

Our concrete suggestions in result reporting are as follows. For whole-brain voxelwise analysis under the conventional framework, we recommend that one adopt a "highlight but not hide" approach. Specifically, it is preferable to highlight brain regions with some extent of statistical evidence (e.g., a cluster threshold of 10 voxels at the voxelwise *p*-value of 0.05) while gradually fading away for the rest (Fig. 1F), instead of the conventional dichotomization methodology (Fig. 1D). Region-level results under the conventional framework can be reported in a table that should include *both* the estimation of effect magnitude *and* the corresponding uncertainty (standard error or uncertainty interval). For region-based analysis under the Bayesian framework, we suggest, if space permits, the adoption of a more informative presentation than a table by showing each full posterior distribution as illustrated in Figs. 6A, 7, and 9.

We note that the information contained in the Bayesian results is much richer and more straightforward in result interpretation. For example, the posterior distributions in Figs. 6A, 7, and 9 show the full range of effect estimates and their uncertainty. In addition, each posterior distribution directly reveals the probability of seeing the effect in any range (e.g., being positive or greater than 0.2) conditional on the current data. As a comparison, even if the analysis under the massively univariate framework is performed at the voxel level, the significance level (e.g., 0.05) adjusted for multiple testing can only be applied to a whole set of spatial blobs, not to individual voxels, due to dichotomization; thus, it would be difficult to attach some sense of uncertainty for each spatial blob.

Our assessment and recommendation regarding modeling and result communication are summarized in Fig. 10. As of 2021, investigators have at their disposal a vast array of tools for the statistical analysis of FMRI data. The majority of them maintain a traditional focus on the conventional way of thinking of inferences in terms of "true" and "false" effects. In this paper, we discussed several problems with applying standard null hypothesis significance testing to FMRI data. We favor a view of neuroimaging effects in terms of a continuum of statistical evidence, with a large number of small effects dominating, instead of islands of strong/true effects that should be discerned from false positives. We propose that Bayesian multilevel modeling has considerable potential in complementing, if not improving, statistical practices in the field, one that emphasizes effect estimation rather than statistical dichotomization, with the goal of "seeing the forest for the trees" and improving the quality and reproducibility of research in the field.

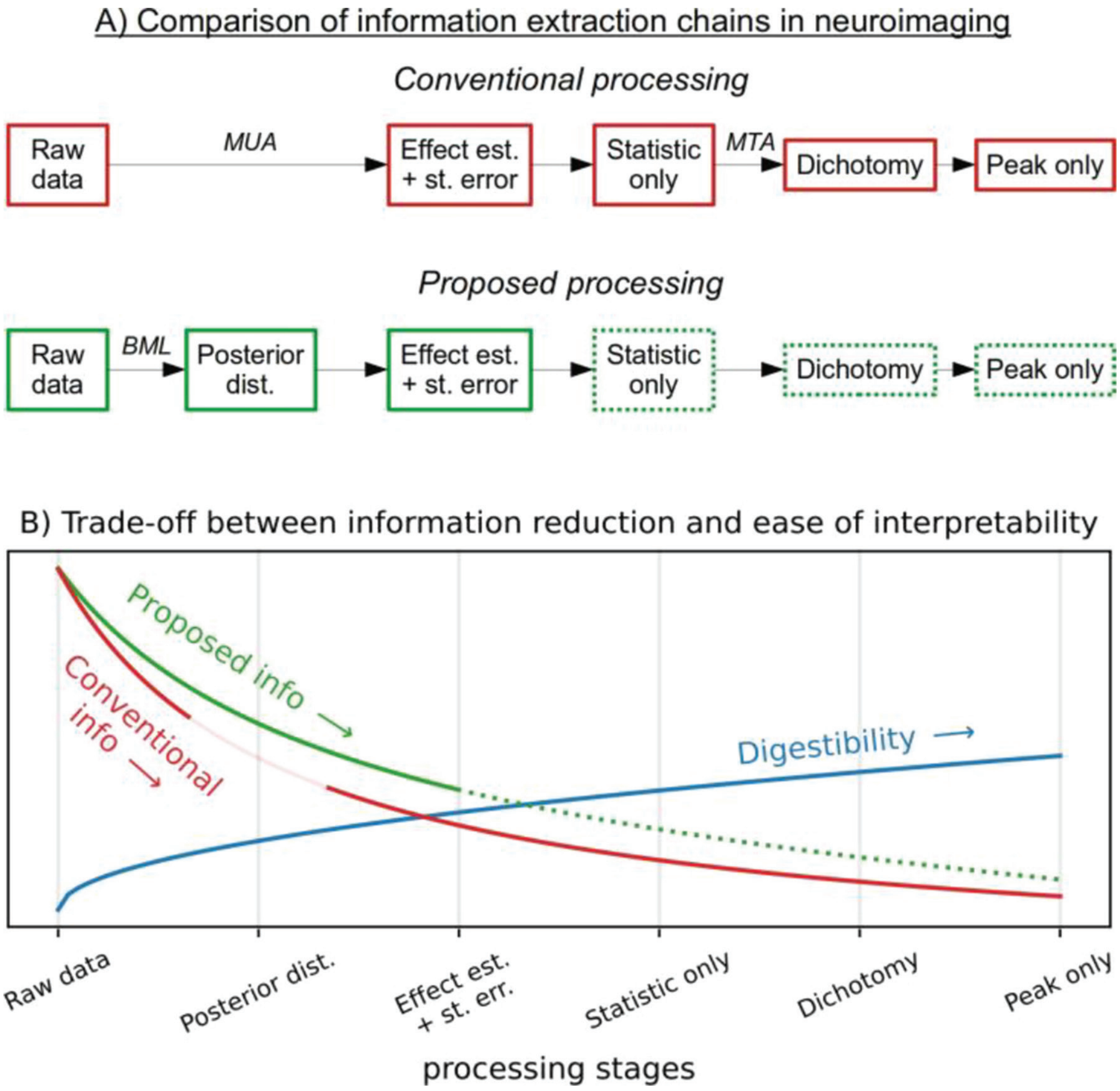


Fig. 10. Comparison of FMRI information extraction for conventional and proposed Bayesian multilevel (BML) approaches (cf. Fig. 2). (A) The two approaches run parallel, but in the "proposed" first step, BML puts data into a single model (removing the need for multiple testing adjustment later), and the information is partially pooled and shared across space. (B) The proposed multilevel framework produces an intermediate output of posterior distributions (lacking in the conventional approach) that carry rich information about parameter and model fitting. Partial pooling also improves model efficiency and avoids potential overfitting. This information advantage over the conventional method carries on to later stages. Thus, while the "digestibility" of results increases similarly at each stage, the drop-off in information content is slower in the proposed approach. The dotted part of the proposed steps reflects that we strongly suggest not including the steps that the traditional approaches at present perform, due to the wasteful information loss incurred.

In neuroimaging, research groups acquire different sized datasets with different sample sizes and paradigms varying to some degree. With various preprocessing and modeling approaches available in the community, some extent of result variation is expected and unavoidable. All these factors contribute to an expected variability in reported results, and it need not be considered inherently problematic. To accurately combine multiple studies and determine the levels of variability present, one would need to make a model using their *un*thresholded results and preferably both their effect estimates and uncertainty information. Otherwise, small outcome differences can appear to be much larger, when passed through the dichotomization sieve. Thus, it is the result presentation (e.g., highlight but not hide, show effect magnitude instead of statistical evidence only, revealing model details, etc.) that would conduce to the convergence of a specific research hypothesis across teams. We believe that the abandoning of result dichotomization is one small step toward reducing variability due to artificial thresholding. We agree with NARPS's suggestion of encouraging original statistical results being submitted to a public site. However, more improvements would be needed. For example, such public results at present are still restricted to statistical evidence without the availability of effect magnitude information. Furthermore, proper presentations in publications remain a crucial interface for direct scientific communication and exchange. Therefore, in repositories such as NeuroVault (65) where researchers are able to upload their study results for community sharing, we recommend that researchers upload their effect estimate and uncertainty data, in addition to (or instead of) just statistical values.

CONCLUSIONS

Three aspects of information waste are involved in the conventional neuroimaging data analysis: (a) the implicit adoption of the principle of insufficient reason in massively univariate analysis, (b) hard dichotomization through multiple testing adjustment, and (c) sole focus on statistical evidence without revealing effect magnitude and the associated uncertainty. Under the Bayesian multilevel framework, the data hierarchy across space can be captured and regularized to prevent overfitting and information waste. In addition, full results are available without artificial dichotomization. For future analyses, one may consider the following three aspects regardless of the modeling framework: (1) incorporate data hierarchies into modeling; (2) avoid hard thresholding; and (3) report results that contain both estimate magnitude and uncertainty.

ACKNOWLEDGMENTS

GC, PAT, and RWC were supported by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA. JS was supported by the National Institute of Mental Health (K23MH113731). LP was supported by the National Institute of Mental Health (MH071589 and MH112517).

REFERENCES

1. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature. 2020 Jun;582(7810):84–88.
2. Zhang L, Guindani M, Versace F, Engelmann JM, Vannucci M. A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. Annals of Applied Statistics. 2016 Jun;10(2):638–66.
3. Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. Journal of Cerebral Blood Flow & Metabolism. 1992 Nov;12(6):900–18.
4. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional Magnetic Resonance Imaging (fMRI): Use of a cluster-size threshold. Magnetic Resonance in Medicine. 1995 May;33(5):636–47.
5. Smith SM, Nichols TE. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009 Jan;44(1):83–98.
6. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences of the United States of America. 2016 Jul;113(28):7900–5.
7. Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. NeuroImage. 2014 May;91:412–9.
8. Chen G, Taylor PA, Cox RW, Pessoa L. Fighting or embracing multiplicity in neuroimaging? Neighborhood leverage versus global calibration. NeuroImage. 2020 Feb;206:116320.
9. Eva B. Principles of indifference. Journal of Philosophy. 2019 Apr;116(7): 390–411.
10. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. Boca Raton: Chapman and Hall/CRC; 2013.
11. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305–7.
12. Scargle J. Publication bias: The "file-drawer" problem in scientific inference. Journal of Scientific Exploration. 2000 Jan;14:91–106.
13. van Zwet EW, Cator EA. The significance filter, the winner's curse and the need to shrink. Statistica Neerlandica. 2021 Mar;75(4):437–52.
14. Gelman A, Carlin J. Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. Per- spectives on Psychological Science. 2014 Nov;9(6):641–51.
15. Nuzzo R. Scientific method: Statistical errors. Nature News. 2014 Feb; 506(7487):150.
16. Wasserstein RL, Lazar NA. The ASA statement on p-values: Context, process, and purpose. The American Statistician. 2016 Apr;70(2):129–33.
17. Gonzalez-Castillo J, Saad ZS, Handwerker DA, Inati SJ, Brenowitz N, Bandettini PA. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proceedings of the National Academy of Sciences. 2012 Apr;109(14):5487–92.
18. Barton N, Hermisson J, Nordborg M. Why structure matters. eLife. 2019 Mar;8:e45380.
19. Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Børglum AD, Breen G, et al. Psychiatric genomics: An update and an agenda. American Journal of Psychiatry. 2017 Oct;175(1):15–27.
20. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics. 2005 Sep;30(3):261–93.
21. Chen G, Saad ZS, Nath AR, Beauchamp MS, Cox RW. FMRI group analysis combining effect estimates and their variances. NeuroImage. 2012 Mar;60(1):747–65.
22. Chen G, Taylor PA, Cox RW. Is the statistic value all we should care about in neuroimaging? NeuroImage. 2017 Feb;147:952–9.
23. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika. 2013 Oct;78(4):685–709.
24. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness. 2012 Apr;5(2):189–211.
25. Chen G, Xiao Y, Taylor PA, Rajendra JK, Riggins T, Geng F, et al. Handling multiplicity in neuroimaging through Bayesian lenses with multilevel modeling. Neuroinformatics. 2019 Oct;17(4):515–45.
26. Xiao Y, Geng F, Riggins T, Chen G, Redcay E. Neural correlates of developing theory of mind competence in early childhood. NeuroImage. 2019 Jan;184:707–16.
27. Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Haller SP. Trial and error: A hierarchical modeling approach to test-retest reliability. NeuroImage. 2021 Dec;245:118647.
28. Chen G, Bürkner PC, Taylor PA, Li Z, Yin L, Glen DR, et al. An integrative Bayesian approach to matrix-based analysis in neuroimaging. Human Brain Mapping. 2019;40(14):4072–90.
29. Lima Portugal LC, Alves RdCS, Junior OF, Sanchez TA, Mocaiber I, Volchan E, et al. Interactions between emotion and action in the brain. NeuroImage. 2020 Jul;214:116728.
30. Kantonen T, Karjalainen T, Isojärvi J, Nuutila P, Tuisku J, Rinne J, et al. Interindividual variability and lateralization of μ-opioid receptors in the human brain. NeuroImage. 2020 Aug;217:116922.
31. Limbachia C, Morrow K, Khibovska A, Meyer C, Padmala S, Pessoa L. Controllability over stressor decreases responses in key threat-related brain areas. bioRxiv. 2020 Jul:2020.07.11.198762.
32. Muller P, Parmigiani G, Rice K. FDR and Bayesian multiple comparisons rules. Johns Hopkins University, Dept of Biostatistics Working Papers. 2006 Jul.
33. Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics. 2010 Oct;38(5):2587–619.
34. Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, et al. Bayesian workflow. arXiv:201101808 [stat]. 2020 Nov.
35. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review. 2016 Feb;23(1):103–23.
36. Gelman A, Tuerlinckx F. Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics. 2000 Sep; 15(3):373–90.
37. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and STAN. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2020.

38. Teng M, Nathoo FS, Johnson TD. Bayesian analysis of functional magnetic resonance imaging data with spatially varying auto-regressive orders. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019;68(3):521–41.
39. Teng M, Johnson TD, Nathoo FS. Time series analysis of fMRI data: Spatial modelling and Bayesian computation. Statistics in Medicine. 2018 Aug;37(18): 2753–70.
40. Yu CH, Prado R, Ombao H, Rowe D. A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fMRI. Journal of the American Statistical Association. 2018 Oct;113(524):1395–410.
41. Yu Z, Prado R, Cramer SC, Quinlan EB, Ombao H. A Bayesian model for activation and connectivity in task-related fMRI data. In: Jeliazkov I, Tobias JL, editors. Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A. Vol. 40A of Advances in Econometrics. Emerald Publishing Limited; 2019. p. 91–132.
42. Bezener M, Eberly LE, Hughes J, Jones G, Musgrove DR. Bayesian spatiotemporal modeling for detecting neuronal activation via functional magnetic resonance imaging. In: Härdle WK, Lu HHS, Shen X, editors. Handbook of Big Data Analytics. Springer Handbooks of Computational Statistics. Cham: Springer International Publishing; 2018. p. 485–501.
43. Ferreira da Silva AR. A Bayesian multilevel model for fMRI data analysis. Computer Methods and Programs in Biomedicine. 2011 Jun;102(3):238–52.
44. Flandin G, Penny WD. Bayesian fMRI data analysis with sparse spatial basis function priors. NeuroImage. 2007 Feb;34(3):1108–25.
45. Penny WD, Trujillo-Barreto NJ, Friston KJ. Bayesian fMRI time series analysis with spatial priors. NeuroImage. 2005 Jan;24(2):350–62.
46. Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J. Classical and Bayesian inference in neuroimaging: Theory. NeuroImage. 2002 Jun;16(2):465–83.
47. Friston KJ, Penny W. Posterior probability maps and SPMs. NeuroImage. 2003 Jul;19(3):1240–9.
48. Penny W, Kiebel S, Friston K. Variational Bayesian inference for fMRI time series. NeuroImage. 2003 Jul;19(3):727–41.
49. Wang G, Muschelli J, Lindquist MA. Moderated t-tests for group-level fMRI analysis. NeuroImage. 2021 Aug;237:118141.
50. Whiteman AS, Bartsch AJ, Kang J, Johnson TD. Bayesian inference for brain activity from functional Magnetic Resonance Imaging collected at two spatial resolutions. arXiv:210313131 [stat]. 2021 Mar.
51. Woolrich MW, Behrens TEJ, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage. 2004 Apr;21(4):1732–47.
52. Helske J, Helske S, Cooper M, Ynnerman A, Besançon L. Can visualization alleviate dichotomous thinking? Effects of visual representations on the cliff effect. IEEE Transactions on Visualization and Computer Graphics. 2021 Aug;27(08):3397–409.
53. Allen E, Erhardt E, Calhoun V. Data visualization in the neurosciences: Overcoming the curse of dimensionality. Neuron. 2012 May;74(4):603–8.
54. Gelman A. Ethics in statistical practice and communication: Five recommendations. Significance. 2018;15(5):40–43.
55. Coleman EB. Generalizing to a language population. Psychological Reports. 1964 Feb;14(1):219–26.
56. Clark HH. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior. 1973 Aug;12(4):335–59.
57. Westfall J, Nichols TE, Yarkoni T. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research. 2017 Mar;1:23.
58. Chen G, Padmala S, Chen Y, Taylor PA, Cox RW, Pessoa L. To pool or not to pool: Can we ignore cross-trial variability in FMRI? NeuroImage. 2021;225(4):117496.
59. Hedge C, Powell G, Sumner P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods. 2018 Jun;50(3):1166–86.
60. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, et al. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychological Science. 2020 Jun;31:792–806.
61. Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Taylor PA, et al. Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. bioRxiv. 2021 Jul:2021.07.15.452548.
62. Rosenblatt JD, Finos L, Weeda WD, Solari A, Goeman JJ. All-resolutions inference for brain imaging. NeuroImage. 2018 Nov;181:786–96.
63. Silva ARFd. cudaBayesreg: Parallel implementation of a Bayesian multilevel model for fMRI data analysis. Journal of Statistical Software. 2011 Oct;44(1):1–24.
64. Češnovar R, Bronder S, Sluga D, Demšar J, Ciglarič T, Talts S, et al. GPU-based parallel computation support for Stan. arXiv:190701063 [cs, stat]. 2020 May.
65. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, et al. NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Frontiers in Neuroinformatics. 2015;9:8.