bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Dichotomous thinking and informational waste in neuroimaging

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, USA
[b]Department of Psychiatry, University of Colorado, USA
[c]Section on Functional Imaging Methods, NIMH, National Institutes of Health, USA
[d]Department of Psychology, Department of Electrical and Computer Engineering, and Maryland Neuroimaging Center, University of Maryland, USA

Abstract

Neuroimaging relies on separate statistical inferences at tens of thousands of spatial locations. Such massively univariate analysis typically requires adjustment for multiple testing in an attempt to maintain the family-wise error rate at a nominal level of 5%. We discuss how this approach is associated with substantial information loss because of an implicit but questionable assumption about the effect distribution across spatial units. To improve inference efficiency, predictive accuracy, and generalizability, we propose a Bayesian multilevel modeling framework. In addition, we make four actionable suggestions to alleviate information waste and to improve reproducibility: (1) abandon strict dichotomization; (2) report full results; (3) quantify effects, and (4) model data hierarchy.

1 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 1-3 mm on each side. As data are collected across time when a person performs a task, 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, false discovery rate, or some other approach.

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*

*Corresponding author. E-mail address: gangchen@mail.nih.gov

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

(informally interpreted as a "true" effect) while the other part is ignored (often misinterpreted as "not true") and by convention omitted and hidden from public view.

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 vs. nonsignificant" maxim. In this context, we address the following questions:

1) Are conventional multiple testing adjustment methods informationally wasteful?
2) The NARPS study suggested that there was "substantial variability" in reported results across teams of investigators studying the same dataset. Is this conclusion dependent, at least in part, on the practice of drawing inferences binarily (i.e., "significant" vs. "non significant")?
3) How can the neuroimaging field improve analysis and reporting practices to improve replicability?

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 dichotomous thinking in neuroimaging, we do not discuss the broader literature on Bayesian methods applied to FMRI[2].

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 data for the effect in question (e.g., task A vs. task B) is unlikely to be observed under the condition of null effect. 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 the activation cluster that would be unlikely to be observed 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 approach[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].

Four limitations are associated with multiple testing adjustment leveraged through spatial extent[8].

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

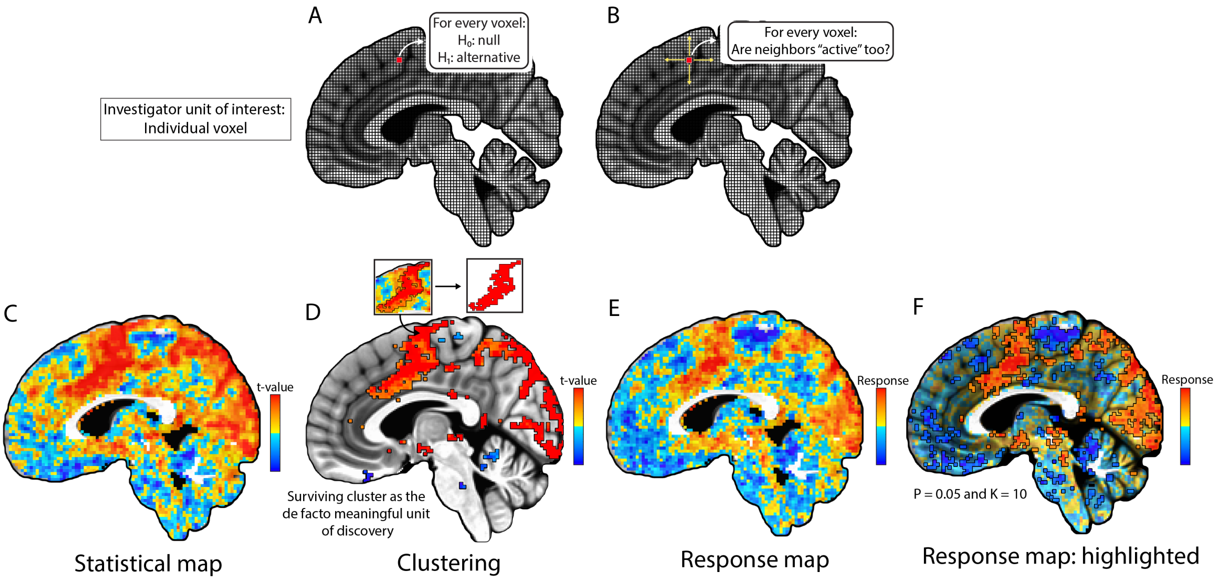


Figure 1: Statistical inferences in neuroimaging. (A) Schematic view of standard analysis: each voxel among tens of thousands of voxels is tested against the null hypothesis (voxel not drawn to scale). (B) Clusters of contiguous voxels with strong statistical evidence are adopted to address the multiple 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: "highlight" parts with strong statistical evidence, but do not "hide" the rest.

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 lose direct interpretability. Although this issue has been discussed in the past[7], it remains underappreciated, and researchers commonly do not adjust their presentations to match the cluster-level effective resolution.

2) *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 strength while smaller regions would have to require much stronger statistical strength. Therefore, multiple testing adjustment always penalizes 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.

3) *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 (e.g., "small volume correction"), resulting in some extent of arbitrariness: one cluster may survive or fail depending on the data volume. Because of this vulnerability, it is not easy to draw a clear line between a justifiable reduction of data and an exploratory search.

4) *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 at the cluster level. Recent effort has been taken to address the issue of assigning uncertainty at the spatial extent level[9]. A cluster either survives or

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

not under a dichotomous decision; on the other hand, it remains challenging to have, for example, a standard error ("error bar") associated with the average effect at the cluster level.

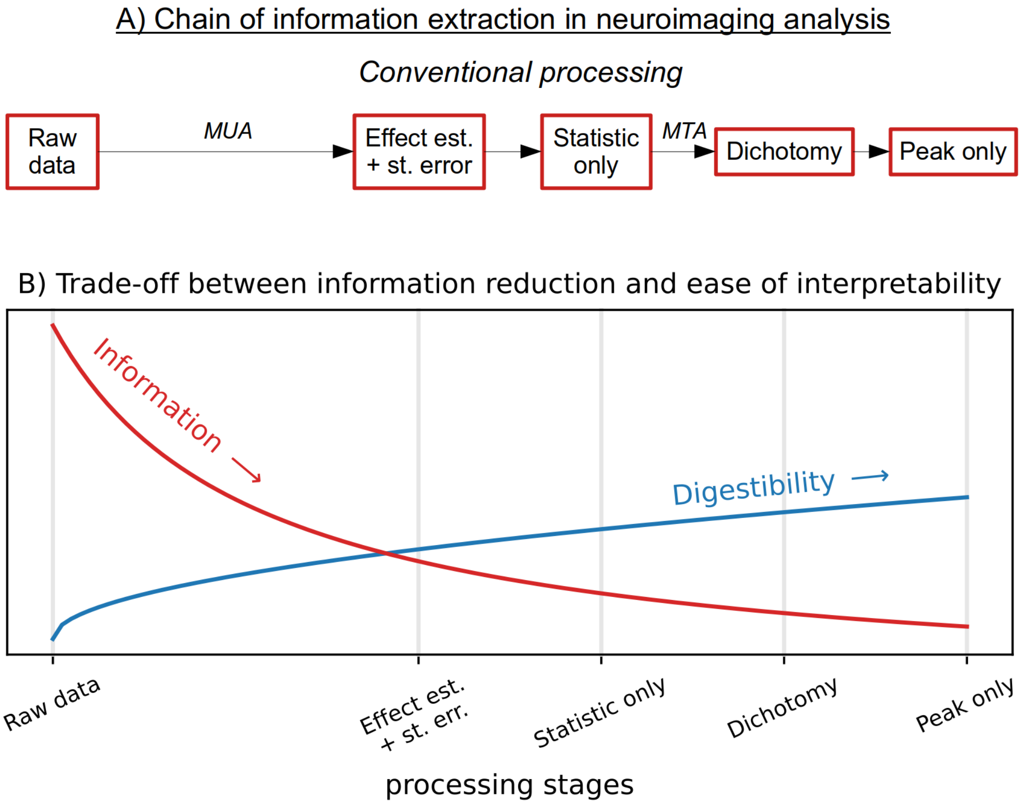


Figure 2: A schematic of conventional information extraction in neuroimaging. (A) The processing chain starts with raw data. Massively univariate analysis (MUA) produces an effect estimate and its uncertainty (standard error) at every spatial unit. These are reduced to a single statistic map, which is then dichotomized using thresholding with multiple testing adjustment (MTA); finally, many studies summarize 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 the non-dichotomized effect estimate and standard errors, if possible, would be preferable, striking a better balance between information loss and interpretability.

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, where as statistical models aim 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. 2 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 effect 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, etc.). "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 then 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, below, we argue that the overall procedure leads to information

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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.

2.1 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 $-\infty$ to $+\infty$ (Fig. 3A), at times discussed as the *principle of indifference* 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.

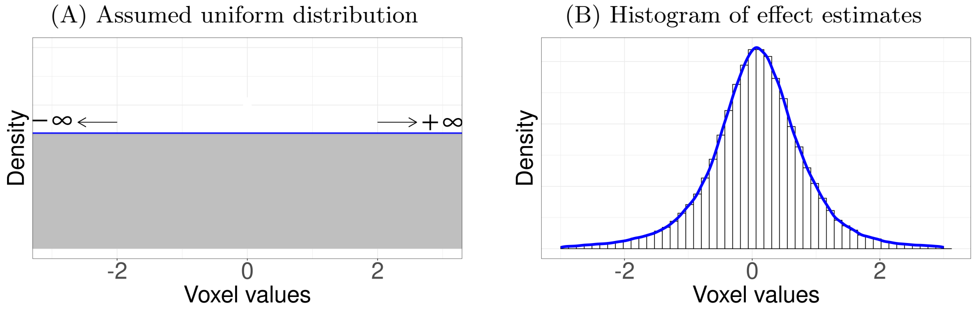


Figure 3: Distributions of effects ("activation strength") across space. (A) Under the conventional massively univariate analysis framework, effects across all spatial units (voxels) are implicitly assumed to be drawn from a uniform distribution. Accordingly, the effect at each spatial unit can assume any value within $(-\infty, +\infty)$ with equal likelihood. (B) Histogram of effect estimates (percent signal change) across 153768 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.

In addition to potentially excessive penalties due to information waste, the principle of indifference has another important ramification: *overfitting*. Under massively univariate analysis, the model is free to fit the voxel's data in any way it can as all effects 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 standard massively univariate approach accurately estimates the effect at the spatial unit level (via least squares or maximum likelihood), the approach tends to fit individual voxels overly close to the sample data at hand, possibly paying 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* (termed "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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

stance of complete ignorance (i.e., uniform distribution).

Technically, we can say that the voxel-effect distribution, $\mathcal{N}(\mu,\ \sigma^2)$, forms a *prior distribution* in the Bayesian sense where both the mean μ and the standard deviation σ are 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 across the spatial units through a process of "information sharing", regularization or *partial pooling*. For example, if most of the individual effects 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 all the studies 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 70 teams simultaneously analyzed the data and provided separate results. We agree that the adoption of a Gaussian prior is a sensible approach: 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 analysis across different analytical pipelines; rather, information integration through a "forest perspective" can be equally applied to modeling across all hierarchies, including voxel, region, and participant levels.

3 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 adjustment 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 multiple testing 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 confusion[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, 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 informa-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

tional 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 significance statements 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 participant samples were considered[17]. More generally, many domains of research appear to be characterized by a very large number of "small effects", as opposed to 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.

We propose that a more productive approach is to refocus research objectives away from trying to uncover "real" effects. Instead, 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, right). 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 this context, even the typical language of "activated voxels/regions" comes with substantial perils; we encourage further discussion about better and more nuanced ways of summarizing research findings.

3.1 Neglect of effect magnitude and uncertainty measures

Statistical significance combines two underlying pieces of information: the effect estimate and its uncertainty (consider the t-statistic, which is the ratio of an effect estimate to its standard error). 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 course based on probability theory. 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 standard 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 the NARPS study. Given the findings from the 70 independent teams, NARPS performed two types of meta analysis: one with binarized team reports (logistic regression), and another solely based on statistical values. In the binarized case, the result of each individual study was considered either present (value of 1) or absent (value of 0). NARPS 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 significantly different from a statistically insignificant one (Fig. 4, left). 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 $p < 0.05$), because they have quantitatively different uncertainties—

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

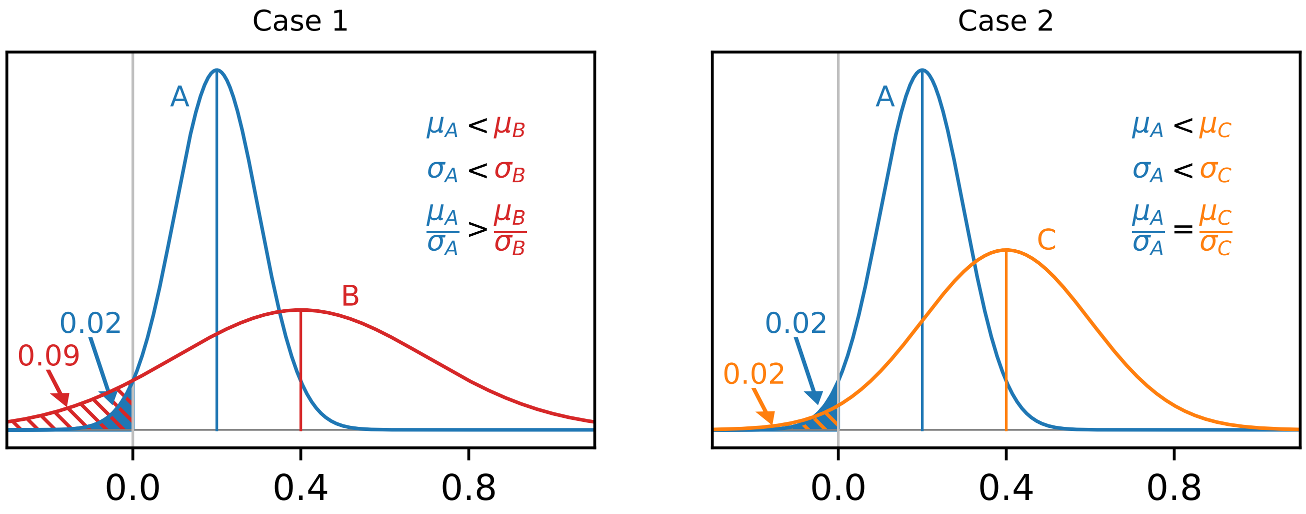


Figure 4: Implications of dichotomization in conventional statistical practice. (Left) What is the difference between a statistically significant result and one that does not cross threshold? Between the two hypothetical effects that follow Gaussian distributions ($\mu_A = 0.2$, $\sigma_A = 0.1$ (blue); $\mu_B = 0.4$, $\sigma_B = 0.3$ (red)), only effect A would be considered statistically significant. However, note that the difference between the two effects is not statistically significant ($p = 0.26$, one-sided), and effect B is mostly larger than A with a probability of 0.74. (Right) How much information is lost due to the focus on binary statistical decisions? The two hypothetical effects (with normal distributions: $\mu_A = 0.2$, $\sigma_A = 0.1$ (blue); $\mu_C = 0.4$, $\sigma_C = 0.2$ (orange)) have the same t- and p-values, and would be deemed indistinguishable in terms of statistical evidence alone. However effect C is mostly larger than effect A with a probability of 0.81. This comparison illustrates the information loss when the sole focus is on statistic or p value, which is further illustrated between the second and third blocks in Fig. 2.

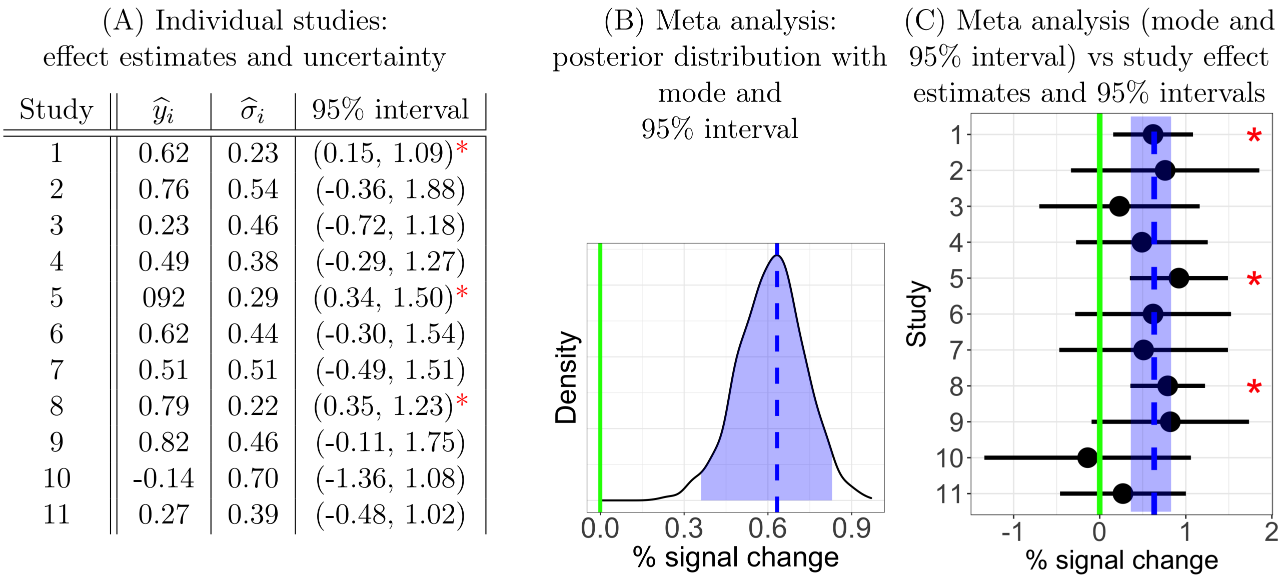
one can think of one having a much wider uncertainty interval than the other, although both exclude zero (Fig. 4, right). Thus, they are treated as providing the same amount of statistical evidence despite potential nontrivial differences in both effect magnitude and uncertainty. These examples illustrate the extent of information loss due to the 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., activation in the amygdala due to fearful and neutral faces), or to the case considered by NARPS in which different teams investigated 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 considers applying multiple testing adjustments to the statistical threshold due to having 11 parallel inferences: with adjustment, none of the studies would survive.

Instead of a logistic regression based on binarized assessments, an integrative meta analysis can be performed by combining the full results: *both* the effect estimate and uncertainty from each study. Let us assume that the effect estimates, $\widehat{y}_i$ ($i = 1, 2, ..., 11$), are normally distributed $\widehat{y}_i \sim \mathcal{N}(\theta_i,\ \widehat{\sigma}_i^2)$, with mean θ_i and variance $\hat{\sigma}_i^2$. In addition, assume that the effects themselves, θ_i, follow a Gaussian distribution $\theta_i \sim \mathcal{N}(\mu,\ \tau^2)$, 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 the overall mean μ (Fig. 5B) with an average effect $\widehat{\mu} = 0.61$ and a 95% uncertainty interval of $[0.34, 0.85]$. When this posterior uncertainty interval is reviewed together with the

[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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



(A) Individual studies: effect estimates and uncertainty

Study	$\widehat{y}_i$	$\widehat{\sigma}_i$	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 study effect estimates and 95% intervals

Figure 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, $\widehat{y}_i$ (where i is the study index), and its standard error, $\widehat{\sigma}_i$. A total of 3/11 effects would be deemed statistically significant (red asterisk) 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 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 $\widehat{y}_i$, horizontal lines show $\widehat{\sigma}_i$, and red asterisks indicate 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.

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 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 estimates 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 multiple testing adjustment. 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 including statistic values is a step in the right direction, it is an insufficient one. Displaying 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

$$\widehat{\mu} = \frac{\displaystyle\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_i^2 + \tau^2}\widehat{y}_i}{\displaystyle\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_i^2 + \tau^2}}, \tag{1}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

study (inversely) contributing to the weight.

The preceding 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 reduction that limits many kinds of subsequent types of data analysis[20]. 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 to 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, right). 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.

Returning to the NARPS investigation, they performed a second meta analysis solely based on statistic values. Under this approach without dichotomization, the findings across teams were substantially more consistent with one another, 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 the 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, the conclusion bias would have been further exacerbated when results were binarized with "statistically nonsignificant" ones unreported and hidden.

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, while reasonable in some statistical settings, in the case of FMRI disregards distributional information concerning effect magnitude across the brain (Fig. 3).
2) Hard thresholding carries with it a fair mount 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 meta analysis and study reproducibility (and makes spotting spurious results less straightforward).

In the next section, we describe how Bayesian multilevel modeling provides a modeling paradigm that can contribute to addressing the issues above.

3.2 Bayesian multilevel modeling

In this section, we briefly describe Bayesian multilevel modeling. We start with building up the structure by first considering simple data y_{ij} ($i = 1, 2, ..., n$; $j = 1, 2, ..., k$) from n subjects that are longitudinally measured under k time points with a predictor x_{ij}, using the form $y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$ with intercept α_i, slopes β_i, and residuals ϵ_{ij}. To appreciate the flexibility of the approach, this model is sometimes referred

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

to as a "varying-intercept/varying-slope" model akin to those commonly adopted in a multilevel framework:

$$\begin{aligned}
y_{ij} &\sim \mathcal{N}(\mu_{ij},\ \sigma_\epsilon^2) \\
\mu_{ij} &= \alpha_i + \beta_i x_{ij} \\
\alpha_i &\sim \mathcal{N}(\alpha,\ \sigma_\alpha^2) \\
\beta_i &\sim \mathcal{N}(\beta,\ \sigma_\beta^2) \\
\alpha &\sim \mathcal{N}(0,\ 1) \\
\beta &\sim \mathcal{N}(0,\ 1) \\
\sigma_\alpha &\sim \text{HalfCauchy}(1) \\
\sigma_\beta &\sim \text{HalfCauchy}(1)\,.
\end{aligned} \tag{2}$$

What makes the model "multilevel" is that it involves the hierarchical levels of subject i and time j. The notation α_i indicates that each subject i has a unique intercept; likewise, the notation β_i indicates that each subject i is given a unique slope. The first line specifies the *likelihood* or the distributional assumption for the data y_i. The expression for μ_i specifies a linear relationship with a single predictor, x (adding more predictors is straightforward). The third and fourth lines are *priors*: the varying intercepts follow a Gaussian distribution with a grand intercept α with standard deviation σ_α; likewise, the varying slopes follow a Gaussian distribution with a grand slope β with standard deviation σ_β. Importantly, the parameters of the prior distributions are learned from the data. Finally, the last four lines specify so-called *hyperpriors*, which can be conveniently weakly informative distributions for the means and variances specified in the priors. Note that the hyperpriors defining the standard deviations are positive only.

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 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 an effect 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 apply partial pooling on the studies through weighting as shown in the 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 partial pooling of the estimates 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 were estimated individually. The information is effectively calibrated across spatial units, and the effect estimates tend to be stabilized even when the data are noisy at a given hierarchy level. Because of this conservative nature, the multilevel model aims to control for errors of incorrect magnitude and sign. Furthermore, adjustment for multiplicity is not needed[22], 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 framework can be effectively employed to analyze FMRI data at the region level[8,23], as well as for matrix-based analysis including time series correlations or white-matter properties[25]. 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 a region-level example. In a Bayesian analysis, the outcome is the posterior distribution

[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[21].

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

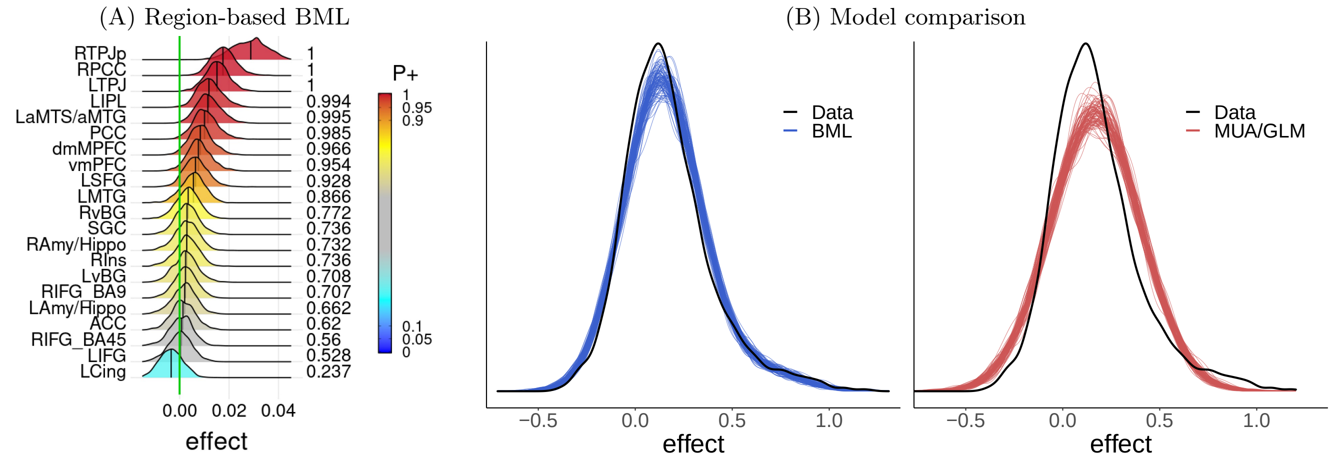


Figure 6: Bayesian multilevel (BML) modeling at the region level. (A) Population-level analysis applied to an FMRI study with 124 subjects[23]. Colors represent values of P^+: the posterior probability that the effect is greater than 0. The analysis revealed that over one third of the regions exhibited considerable statistical evidence in favor of a positive effect. In contrast, with the conventional massively univariate analysis, only two regions survived multiple testing adjustment[24]. (B) BML performance can be assessed and compared to the conventional approach. Posterior predictive checks graphically compare model predictions against raw data. The BML model generated a better fit to the data compared to the general linear model (GLM) employed in the massively univariate analysis (MUA).

that characterize the probability of observing an effect value in a range given the data. Fig. 6A illustrates the results of a recent application at the level of regions[23]. For each region, there is a full posterior distribution that conveys the effect uncertainty, and here we are interested in how much of the area under the curve is to the right/left of zero (green line; the color of the distribution reflects that area). 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 (L SFG) was 0.92, which may be noteworthy in the research context in question. In particular, model fits can be qualitatively assessed by plotting model predictions 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[26]. 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 they were matched in terms of the shocks experienced. As 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. In the standard approach, one would proceed with voxelwise inferential tests (say, a t-test comparing the two groups) followed by a threshold adjustment based on spatial extent to control for multiple testing.

In contrast, the multilevel approach specifies a single model, which combines all data according to natural hierarchical levels of the data. In this particular study, one natural level was that of the participant pair 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 described functionally and anatomically. Accordingly, we subdivided the insula in each hemisphere into around 10 subregions, each of which with approximately 100 voxels. Thus, the subregions comprised another level of the hierarchy. At the most basic level of the hierarchical structure, the unit was the voxel itself.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

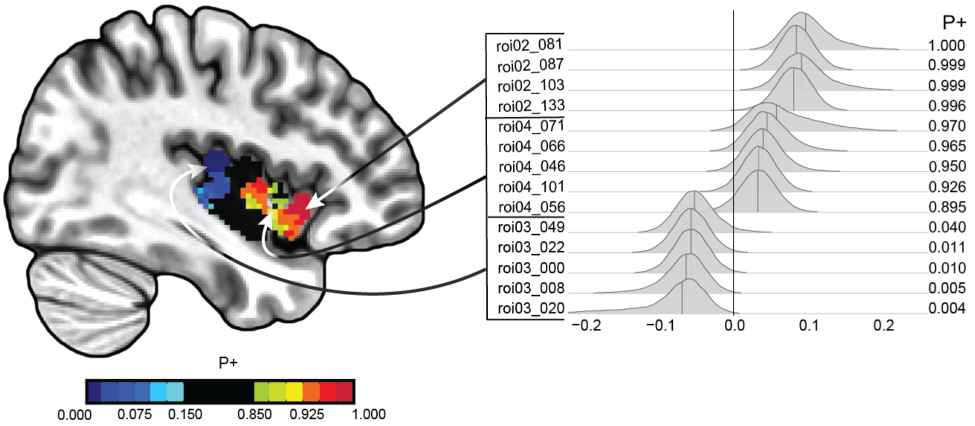


Figure 7: Bayesian multilevel voxelwise results. The right part of the figure illustrates posterior distributions of voxels from three subregions of the insula (voxels selected to illustrate some of the range of statistical evidence). Colors represent values of $P+$: the posterior probability that one 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 following model was employed for the voxel-level data,

$$\Delta_{p,r,v} \sim \mathcal{N}(\mu_{p,r,v},\ \sigma_\epsilon^2)$$
$$\mu_{p,r,v} = \alpha + \beta_p + \gamma_r + \theta_v,$$

where the difference Δ in FMRI responses to shock between a participant pair p in a voxel v belonging to region r is assumed to originate from a Gaussian distribution centered on $\mu_{p,r,v}$ with variance σ_ϵ^2. The second line specifies the response difference 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 modeled by priors as in model (2) (further specifications omitted here for brevity). In this sense, they all 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 model them in terms of varying slopes (thus slopes can vary across regions, for example). This Bayesian machinery allows us to estimate the contributions of participant pairs, regions, and voxels based on the data, the likelihood, and the prior distributions. In the present study, our goal was to understand voxelwise effects (Fig. 7).

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

1) *Handling multiplicity.* The Bayesian approach 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 not needed[22], avoiding excessive penalty due to information waste. We note that some statisticians have suggested other forms of adjustment based on decision theory[2,27,28].

2) *No penalty against small regions.* Under massively univariate analysis, 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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 posteri 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 single integrative framework, the information is shared 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. Thus, the impact on the same spatial unit is relatively negligible even when the total amount of data changes[8].

4) *Model quality control.* Model accuracy and adequacy can be assessed through posterior predictive checks and cross-validations. More generally, the Bayesian approach welcomes an integrated view of the modeling workflow with an iterative process of model development and refinement[29].

5) *Enhanced intepretability.* The Bayesian approach enhances interpretability of analytical results. 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[30]. 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 range. Naturally, parameter values surrounding the peak of the posterior distribution are more likely than those at the extremes (Fig. 5B).

6) *Error controllability.* Instead of the false positive and false negative errors associated with the conventional null hypothesis framework, the Bayesian multilevel framework can be used to control two different errors: *type M* (over- or under-estimation of effect magnitude) and *type S* (incorrect sign)[31].

3.3 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, rather than the p-value (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", and encourages substituting a mechanical rule by 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 the processing of threat. This region and the central nucleus of the amygdala are frequently conceptualized as part of a functional system called the "extended

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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 cut-offs, 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 meta analysis, as discussed in Section 3.1. 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 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[32]. 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[33]. 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[34] (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"[35] in the results of any particular study.

3.4 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 vs. random effects terminology, as opposed to participants, which are treated as random and sampled from a hypothetical population. Technically, this means that researchers cannot generalize beyond the stimuli employed in the experiment (say, the 20 faces used from a given dataset), as recognized several decades ago[36,37]. By modeling trials as instantiations of an idealized condition, a study can generalize the results to trials beyond the confine of those employed in the experiment[38,39]. 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, which tries to capture the mean response, does

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

a reasonable job at explaining signal fluctuations. However, the fit is clearly poor in several places (Fig. 8c). Whereas traditional models in neuroimaging ignore this variability across trials, we propose to explicitly account for it in the underlying statistical model[38,39].

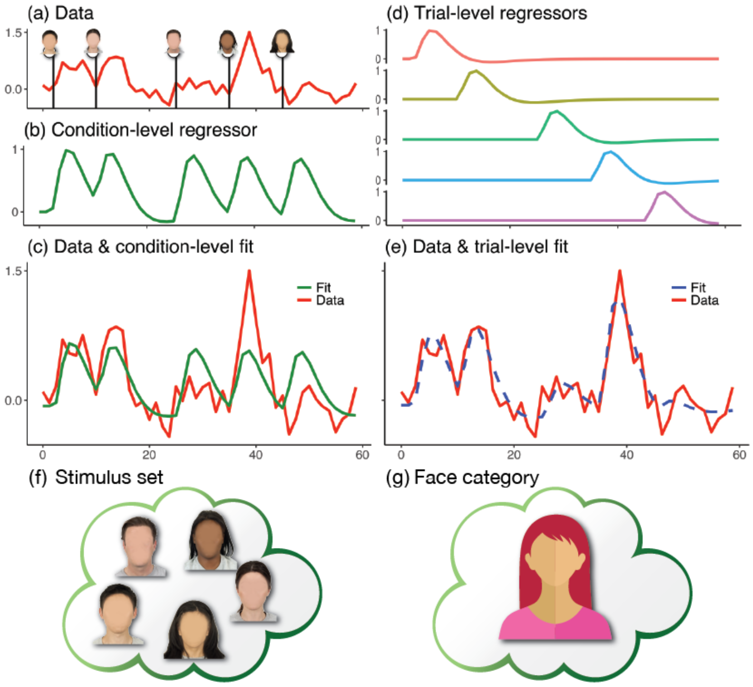


Figure 8: Time series modeling and trial-based analysis. Consider an experiment with five face stimuli. (a) Hypothetical times series. (b) The conventional modeling approach assumes that all stimuli produce the same response, so one regressor is employed. (c) Condition-level effect (e.g., in percent 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.

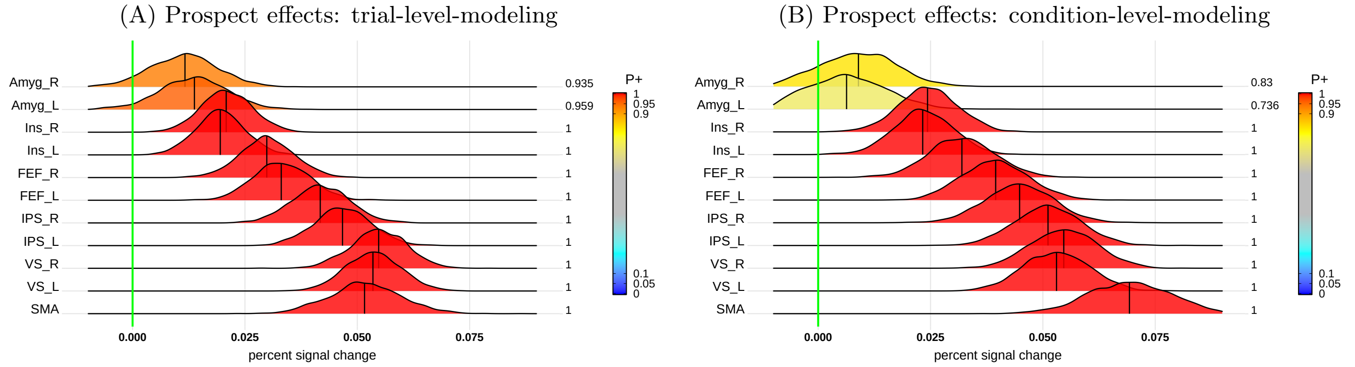


Figure 9: Trial-level versus condition-level modeling. Posterior distributions for the effect of reward (vs. control) cues for each region of interest. Although the two approaches provided comparable results, trial-level modeling (A) showed stronger evidence for left and right amygdala than the condition-level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

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[39]. Indeed, we observed considerable trial-by-trial variability and notable inferential differences when trials were explicitly modeled. For example, as the experiment 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 condition-level modeling.

Trial-level modeling also improves the estimation of test-retest reliability (i.e., the degree of agreement or consistency between measurements carried out under the same conditions). Recent reports have suggested that the test-retest reliability for psychometric[40] and neuroimaging[41] data is rather low when evaluated via

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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 hierarchy of the data structure down to the trial level, which provides a formulation of test-retest reliability that is disentangled from trial-level variability[42]. 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.

4 Conclusions

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. The dream of "whole-brain noninvasive" imaging came 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 developing statistical methods.

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 approach is inefficient and wastes data, even before getting to questions of dichotomization: the initial uniform distribution is far from approximating any realistic brain effect, and the p-values provide 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 data present.

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. Firstly, the Bayesian multilevel modeling approach replaces the massively univariate analysis and the principle of insufficient reason with a single integrated model and 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. If one wanted to, it would be possible to carry on with further collapsing this information into purely statistical form and then dichotomizing, but as noted above, we believe that wastes further information unnecessarily, reduces quality control checks and makes accurate meta analyses difficult.

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 the present paper, we discussed several problems with

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

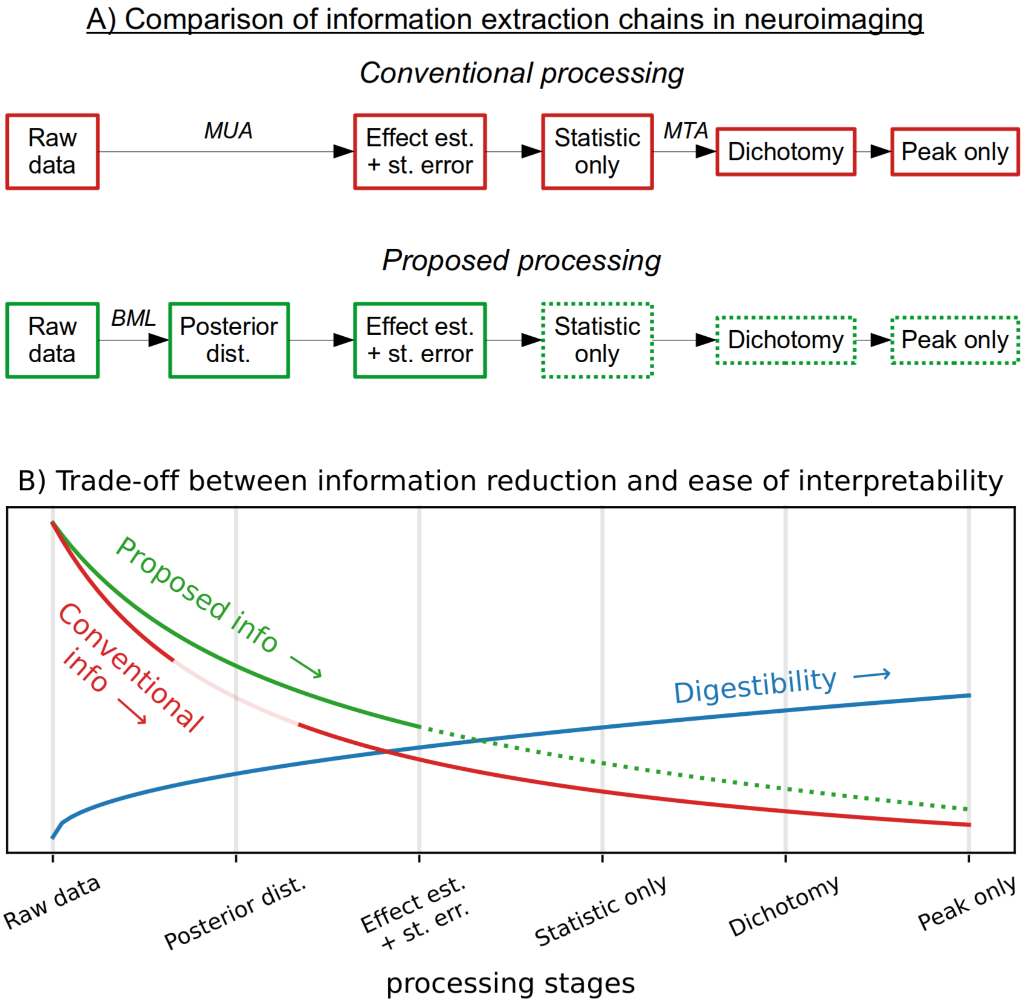


Figure 10: Comparison of FMRI information extraction for conventional and proposed Bayesian multilevel (BML) approaches (cf. Fig. 2). (A) The two approaches run parallel, but in the "proposed" first step BML puts data into a single model (removing the need for multiple testing adjustment later), and 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); these carry rich information about parameter and model fitting. The information pooling also produces effect estimates that retain more information and avoid 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 many traditional approaches at present perform, due to the wasteful information loss incurred.

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.

In neuroimaging, research groups acquire different sized data sets with different subjects 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 preferable both their effect estimates and uncertainty information. Otherwise, small outcome differences can appear to be much larger, when passed through the dichotomization sieve.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

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[43] 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 datasets. For future analyses (quite generally), one may consider the following three aspects: 1) avoid hard thresholding; 2) report both effect estimates and their uncertainties; 3) incorporate the data hierarchy into modeling.

5 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. Number: 7810 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41586-020-2314-9.

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):638–666. Publisher: Institute of Mathematical Statistics. Available from: https://projecteuclid.org/euclid.aoas/1469199888.

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):900–918. Publisher: SAGE Publications Ltd STM. Available from: https://doi.org/10.1038/jcbfm.1992.127.

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;33(5):636–647. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.1910330508. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910330508.

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. Available from: http://www.sciencedirect.com/science/article/pii/S1053811908002978.

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):7900–7905. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4948312/.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

7. Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. NeuroImage. 2014 May;91:412–419. Available from: https://www.sciencedirect.com/science/article/pii/S1053811914000020.

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. Available from: http://www.sciencedirect.com/science/article/pii/S1053811919309115.

9. Bowring A, Telschow FJE, Schwartzman A, Nichols TE. Confidence Sets for Cohen's d effect size images. NeuroImage. 2021 Feb;226:117477. Available from: https://www.sciencedirect.com/science/article/pii/S1053811920309629.

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):305–307. Number: 7748 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/d41586-019-00857-9.

12. Scargle J. Publication bias: the "File-Drawer" problem in scientific inference. Journal of Scientific Exploration. 2000 Jan;14:91–106.

13. Zwet EWv, Cator EA. The Significance Filter, the Winner's Curse and the Need to Shrink. Statistica Neerlandica;n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/stan.12241 tex.ids= vanzwetSignificanceFilterWinner2020, vanzwetSignificanceFilterWinner2020a, vanzwetSignificanceFilterWinner2020b arXiv: 2009.09440. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/stan.12241.

14. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science. 2014 Nov;9(6):641–651. Publisher: SAGE Publications Inc. Available from: https://doi.org/10.1177/1745691614551642.

15. Nuzzo R. Scientific method: Statistical errors. Nature News. 2014 Feb;506(7487):150. Section: News Feature. Available from: http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

16. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016 Apr;70(2):129–133. Publisher: Taylor & Francis. Available from: https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

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):5487–5492. Tex.ids= gonzalezcastilloWholebrainTimelockedActivation2012 ISBN: 9781121049109 publisher: National Academy of Sciences section: Biological Sciences. Available from: https://www.pnas.org/content/109/14/5487.

18. Barton N, Hermisson J, Nordborg M. Why structure matters. eLife. 2019 Mar;8:e45380. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.45380.

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. Publisher: American Psychiatric Publishing. Available from: https://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2017.17030283.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

20. Chen G, Taylor PA, Cox RW. Is the statistic value all we should care about in neuroimaging? NeuroImage. 2017 Feb;147:952–959. Available from: http://www.sciencedirect.com/science/article/pii/S1053811916305432.

21. 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. Available from: https://doi.org/10.1007/s11336-013-9328-2.

22. 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. Publisher: Routledge _eprint: https://doi.org/10.1080/19345747.2011.618213. Available from: https://doi.org/10.1080/19345747.2011.618213.

23. 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):515–545. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6635105/.

24. Xiao Y, Geng F, Riggins T, Chen G, Redcay E. Neural correlates of developing theory of mind competence in early childhood. NeuroImage. 2019 Jan;184:707–716. Tex.ids= xiaoNeuralCorrelatesDeveloping2019. Available from: http://www.sciencedirect.com/science/article/pii/S1053811918319517.

25. 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):4072–4090. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24686.

26. 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. Publisher: Cold Spring Harbor Laboratory Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2020.07.11.198762v1.

27. Muller P, Parmigiani G, Rice K. FDR and Bayesian Multiple Comparisons Rules. Johns Hopkins University, Dept of Biostatistics Working Papers. 2006 Jul. Available from: https://biostats.bepress.com/jhubiostat/paper115.

28. Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics. 2010 Oct;38(5):2587–2619. Publisher: Institute of Mathematical Statistics. Available from: https://projecteuclid.org/euclid.aos/1278861454.

29. Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, et al. Bayesian Workflow. arXiv:201101808 [stat]. 2020 Nov. ArXiv: 2011.01808. Available from: http://arxiv.org/abs/2011.01808.

30. 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):103–123. Available from: https://doi.org/10.3758/s13423-015-0947-8.

31. Gelman A, Tuerlinckx F. Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics. 2000 Sep;15(3):373–390. Available from: https://doi.org/10.1007/s001800000040.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.09.443246; this version posted May 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

32. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and STAN. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2020.

33. Helske J, Helske S, Cooper M, Ynnerman A, Besancon L. Can visualization alleviate dichotomous thinking Effects of visual representations on the cliff effect. IEEE Transactions on Visualization and Computer Graphics. 2021 Apr;(01):1–1. Publisher: IEEE Computer Society. Available from: https://www.computer.org/csdl/journal/tg/5555/01/09405484/1sP1gmcPi7u.

34. Allen E, Erhardt E, Calhoun V. Data Visualization in the Neurosciences: Overcoming the Curse of Dimensionality. Neuron. 2012 May;74(4):603–608. Available from: http://www.sciencedirect.com/science/article/pii/S089662731200428X.

35. Gelman A. Ethics in statistical practice and communication: Five recommendations. Significance. 2018;15(5):40–43. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2018.01193.x. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2018.01193.x.

36. Coleman EB. Generalizing to a Language Population. Psychological Reports. 1964 Feb;14(1):219–226. Publisher: SAGE Publications Inc. Available from: https://doi.org/10.2466/pr0.1964.14.1.219.

37. 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):335–359. Available from: https://www.sciencedirect.com/science/article/pii/S0022537173800143.

38. Westfall J, Nichols TE, Yarkoni T. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research. 2017 Mar;1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5428747/.

39. 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. 2020 Oct:117496. Available from: http://www.sciencedirect.com/science/article/pii/S1053811920309812.

40. 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):1166–1186. Available from: https://doi.org/10.3758/s13428-017-0935-1.

41. 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. Publisher: SAGE PublicationsSage CA: Los Angeles, CA. Available from: https://journals.sagepub.com/doi/10.1177/0956797620916786.

42. Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Haller SP. Beyond the intraclass correlation: A hierarchical modeling approach to test-retest assessment. bioRxiv. 2021 Jan:2021.01.04.425305. Publisher: Cold Spring Harbor Laboratory Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2021.01.04.425305v1.

43. 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. Publisher: Frontiers. Available from: https://www.frontiersin.org/articles/10.3389/fninf.2015.00008/full.