frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fninf.2021.738342

Providing Evidence for the Null Hypothesis in Functional Magnetic Resonance Imaging Using Group-Level Bayesian Inference

Ruslan Masharipov, Irina Knyazeva, Yaroslav Nikolaev, Alexander Korotkov, Michael Didur, Denis Cherednichenko and Maxim Kireev*

N. P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences, Saint Petersburg, Russia

Classical null hypothesis significance testing is limited to the rejection of the point-null hypothesis; it does not allow the interpretation of non-significant results. This leads to a bias against the null hypothesis. Herein, we discuss statistical approaches to 'null effect' assessment focusing on the Bayesian parameter inference (BPI). Although Bayesian methods have been theoretically elaborated and implemented in common neuroimaging software packages, they are not widely used for 'null effect' assessment. BPI considers the posterior probability of finding the effect within or outside the region of practical equivalence to the null value. It can be used to find both 'activated/deactivated' and 'not activated' voxels or to indicate that the obtained data are not sufficient using a single decision rule. It also allows to evaluate the data as the sample size increases and decide to stop the experiment if the obtained data are sufficient to make a confident inference. To demonstrate the advantages of using BPI for fMRI data group analysis, we compare it with classical null hypothesis significance testing on empirical data. We also use simulated data to show how BPI performs under different effect sizes, noise levels, noise distributions and sample sizes. Finally, we consider the problem of defining the region of practical equivalence for BPI and discuss possible applications of BPI in fMRI studies. To facilitate 'null effect' assessment for fMRI practitioners, we provide Statistical Parametric Mapping 12 based toolbox for Bayesian inference.

Keywords: null results, fMRI, Bayesian analyses, human brain, statistical parametric mapping

Edited by:
Ting Zhao,
Janelia Research Campus,
United States

Reviewed by:
Marc De Kamps,
University of Leeds, United Kingdom
Yikang Liu,
United Imaging Intelligence,
United States

Correspondence:
Maxim Kireev
kireev@ihb.spb.ru

Received: *08 July 2021*
Accepted: *05 November 2021*
Published: *02 December 2021*

Citation:
Masharipov R, Knyazeva I, Nikolaev Y, Korotkov A, Didur M, Cherednichenko D and Kireev M (2021) Providing Evidence for the Null Hypothesis in Functional Magnetic Resonance Imaging Using Group-Level Bayesian Inference. Front. Neuroinform. 15:738342. doi: 10.3389/fninf.2021.738342

INTRODUCTION

In the neuroimaging field, it is a common practice to identify statistically significant differences in local brain activity using the general linear model approach for mass-univariate null hypothesis significance testing (NHST) (Friston et al., 1994). NHST considers the probability of obtaining the observed data, or more extreme data, given that the null hypothesis of no difference is true. This probability, or p-value, of 0.01, means that, on average, in one out of 100 'hypothetical' replications of the experiment, we find a difference no less than the one found under the null hypothesis. We conventionally suppose that this is unlikely, therefore, we 'reject the null'; that is, NHST employs 'proof by contradiction' (Cohen, 1994). Conversely, when the p-value is large, it is tempting to 'accept the null.' However, the absence of evidence is not evidence of absence (Altman and Bland, 1995). Using NHST, we can only state that we have 'failed to reject the null.' Therefore, in the classical NHST framework, the question of interpreting non-significant results remains.

The most pervasive misinterpretation of non-significant results is that they provide evidence for the null hypothesis that there is no difference, or 'no effect' (Nickerson, 2000; Greenland et al., 2016; Wasserstein and Lazar, 2016). In fact, non-significant results can be obtained in two cases (Dienes, 2014): (1) the data are insufficient to distinguish the alternative from the null hypothesis, or (2) an effect is indeed null or trivial. To date, the extent to which the problem of making 'no effect' conclusions from non-significant results have affected the field of neuroimaging remains unclear, particularly in functional magnetic resonance imaging (fMRI) studies[1]. Regarding other fields of science such as psychology, neuropsychology, and biology, it was found that in 38–72% of surveyed articles, the null hypothesis was accepted based on non-significant results only (Finch et al., 2001; Schatz et al., 2005; Fidler et al., 2006; Hoekstra et al., 2006; Aczel et al., 2018).

Not mentioning non-significant results at all is another problem. Firstly, some authors may consider non-significant results disappointing or not worth publishing. Secondly, papers with non-significant results are less likely to be published. This publishing bias is also known as the 'file-drawer problem' (Rosenthal, 1979; Ioannidis et al., 2014; de Winter and Dodou, 2015; for evidence in fMRI studies, see Jennings and Van Horn, 2012; Acar et al., 2018; David et al., 2018; Samartsidis et al., 2020). Prejudice against the null hypothesis systematically biases our knowledge of true effects (Greenwald, 1975).

This problem is further compounded by the fact that NHST is usually based on the point-null hypothesis, that is, the hypothesis that the effect is *exactly* zero. However, the probability thereof is zero (Meehl, 1967; Friston et al., 2002a). This means that studies with a sufficiently large sample size will find statistically significant differences even when the effect is trivial or has no *practical* significance (Cohen, 1965, 1994; Serlin and Lapsley, 1985; Kirk, 1996).

Having the means to assess non-significant results would mitigate these problems. To this end, two main alternatives are available: Firstly, there are frequentist approaches that shift from point-null to interval-null hypothesis testing, for example, equivalence testing based on the two one-sided tests (TOST) procedure (Schuirmann, 1987; Wellek, 2010). Secondly, Bayesian approaches that are based on posterior parameter distributions (Lindley, 1965; Greenwald, 1975; Kruschke, 2010) and Bayes factors (Jeffreys, 1939/1948; Kass and Raftery, 1995; Rouder et al., 2009). The advantage of frequentist approaches is that they do not require a substantial paradigm shift (Lakens, 2017; Campbell and Gustafson, 2018). However, it has been argued that Bayesian approaches may be more natural and straightforward than frequentist approaches (Edwards et al., 1963; Lindley, 1975; Friston et al., 2002a; Wagenmakers, 2007; Rouder et al., 2009; Dienes, 2014; Kruschke and Liddell, 2017b). It has long been noted that we tend to perceive lower p-values as stronger evidence for the alternative hypothesis, and higher p-values as evidence for the null, i.e., the 'inverse probability' fallacy as it is referred to by Cohen (1994). This is what we obtain in Bayesian approaches by calculating posterior probabilities. Instead of considering infinite 'hypothetical' replications and employing probabilistic 'proof by contradiction,' Bayesian approaches directly provide evidence for the null and alternative hypotheses given the data, updating our prior beliefs in light of new relevant information. Bayesian inference allows us to 'reject and accept' the null hypothesis on an equal footing. Moreover, it allows us to talk about 'low confidence,' indicating the need to either accumulate more data or revise the study design (see **Figure 1**).

Despite the importance of this issue, and the high level of theoretical elaboration and implementation of Bayesian methods in common neuroimaging software programs, for example, Statistical Parametric Mapping 12 (SPM12) and FMRIB's Software Library (FSL), to date, only a few fMRI studies implemented the Bayesian inference to assess 'null effects' (for example, see subject-level analysis in Magerkurth et al., 2015, group-level analysis in Dandolo and Schwabe, 2019; Feng et al., 2019). Therefore, this study is intended to introduce fMRI practitioners to the methods for assessing 'null effects.' In particular, we focus on Bayesian parameter inference (Friston and Penny, 2003; Penny and Ridgway, 2013), as implemented in SPM12. Although Bayesian methods have been described elsewhere, the distinguishing feature of this study is that we aim to demonstrate the practical implementation of Bayesian inference to the assessment of 'null effects,' and reemphasize its contributions over and above those of classical NHST. We deliberately aim to avoid mathematical details, which can be found elsewhere (Genovese, 2000; Friston et al., 2002a, 2007; Friston and Penny, 2003; Penny et al., 2003, 2005, 2007; Penny and Ridgway, 2013; Woolrich et al., 2004). Firstly, we briefly review the frequentist and Bayesian approaches for the assessment of the 'null effects.' Next, we compare the classical NHST and Bayesian parameter inference using the Human Connectome Project (HCP) and the UCLA Consortium for Neuropsychiatric Phenomics datasets, focusing on group-level analysis. We then consider the choice of the threshold of the effect size for Bayesian parameter inference and estimate the typical effect sizes in different fMRI task designs. To demonstrate how the common sources of variability in empirical data influence NHST and Bayesian parameter inference, we examined their behavior for different sample sizes and spatial smoothing. We also used simulated data to assess BPI performance under different effect sizes, noise levels, noise distributions and sample sizes. Finally, we discuss practical research and clinical applications of Bayesian inference.

THEORY

In this section, we briefly describe the classical NHST framework and review statistical methods which can be used to assess the 'null effect.' We also considered two historical trends

[1] Here are some examples of 'no effect' conclusions that can be found in the fMRI literature: (a) brain area was not activated, (b) brain area was not involved in the function, (c) no effect was found in the brain area ($p > 0.05$), (d) both groups showed no differences, which can be interpreted as evidence against the alternative hypothesis; (e) patients have similar responses to both conditions ($p > 0.05$), that is, they have difficulties in differentiating these conditions; (f) lack of significant correlation during treatment suggest a protective impact of the therapy on brain areas.

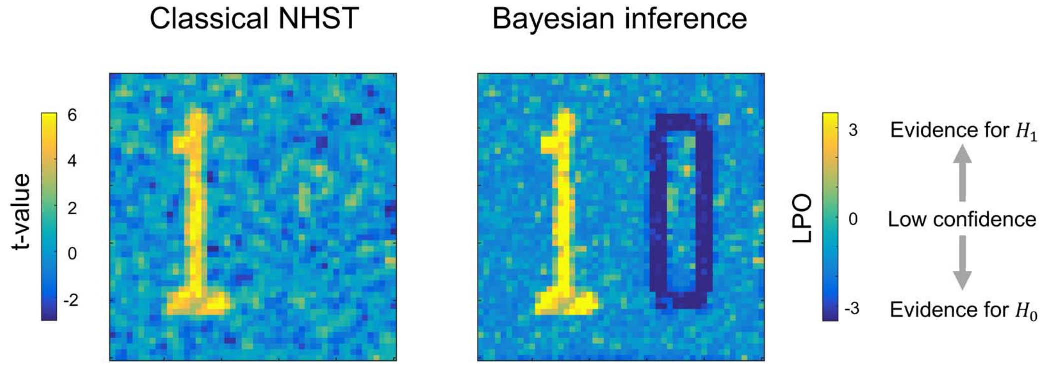

FIGURE 1 | Possible results for the same data, obtained using classical NHST and Bayesian parameter inference. Classical NHST detects only areas with a statistically significant difference ('number one'). Bayesian parameter inference based on the logarithm of posterior probability odds (*LPO*) provides us with additional information that is not available in classical NHST: (1) it provides relative evidence for the null (H_0) and alternative (H_1) hypotheses, (2) it detects areas with a trivial effect size ('number zero'), (3) it indicates 'low confidence' areas surrounding the 'number one' and 'number zero.' To make this conceptual illustration, we generated 100 images consisted of 50 × 50 voxels smoothed by 2 voxel full width at half maximum (FWHM) Gaussian kernel. Data were drawn from normal distributions with different mean, m, and standard deviation, SD. For the 'number one,' m = 0.1, SD = 0.37. For the 'number zero,' m = 0, SD = 0.6. For the 'low confidence' area, m = 0.01, SD = 0.37. *LPOs* were calculated using an effect size threshold of 0.02. The code to recreate the illustration is available online https://github.com/Masharipov/BPI_2021/tree/main/conceptual_illustration.

in statistical analysis: the shift from point-null hypothesis testing to interval estimation and interval-null hypothesis testing (Murphy and Myors, 2004; Wellek, 2010; Cumming, 2013), and the shift from frequentist to Bayesian approaches (Kruschke and Liddell, 2017b).

Classical Null Hypothesis Significance Testing Framework

Most task-based fMRI studies rely on the general linear model approach (Friston et al., 1994; Poline and Brett, 2012). It provides a simple way to separate blood-oxygenated-level dependent (BOLD) signals associated with particular task conditions from nuisance signals and residual noise when analyzing single-subject data (subject-level analysis). At the same time, it allows us to analyze mean BOLD signals within one group of subjects or between different groups (group-level analysis). Firstly, we must specify a general linear model and estimate its parameters:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y are the data (further, D), X is the design matrix, which includes regressors of interest and nuisance regressors, β are the model parameters ('beta values'), and ε is residual noise or error, which is assumed to have a zero-mean normal distribution. At the subject level of analysis, the data are BOLD-signals. At the group level, the data are linear contrasts of parameters estimated at the subject level, which typically reflect individual subject amplitudes of BOLD responses evoked in particular task conditions. In turn, the parameters of the group-level general linear model reflect the group mean BOLD responses evoked in particular task conditions and groups of subjects. The linear contrast of these parameters, $\theta = c\beta$, represents the experimental effect of interest (hereinafter '*the effect*'), expressed as the difference between conditions or groups of subjects.

Next, we test the effect against the point-null hypothesis, H_0: $\theta = \gamma$ (usually, $\theta = 0$). To do this, we use test statistics that summarize the data in a single value, for example, the t-value. For the one-sample case, the t-value is the ratio of the discrepancy of the estimated effect from the hypothetical null value to its standard error. Finally, we calculate the probability of obtaining the observed t-value or a more extreme value, given that the null hypothesis is true (p-value). This is also commonly formulated as the probability of obtaining the observed data or more extreme data, given that the null hypothesis is true (Cohen, 1994). It can be simply written as a conditional probability $P(D+|H_0)$, where '$D+$' denotes the observed data or more extreme data which can be obtained in infinite 'hypothetical' replications under the null (Schneider, 2014, 2018). If this probability is lower than some conventional threshold, or alpha level (for example, $\alpha = 0.05$), then we can 'reject the null hypothesis' and state that we found a statistically significant effect. When this procedure is repeated for a massive number of voxels, it is referred to as 'mass-univariate analysis.' However, if we consider m = 100 000 voxels with no true effect and repeat significance testing for each voxel at $\alpha = 0.05$, we would expect to obtain 5000 false rejections of the null hypothesis (false positives). To control the number of false positives, we must reduce the alpha level for each significance test by applying the multiple comparison correction (Genovese et al., 2002; Nichols and Hayasaka, 2003; Nichols, 2012).

To date, the classical NHST has been the most widely used statistical inference method in neuroscience, psychology, and biomedicine (Szucs and Ioannidis, 2017, 2020; Ioannidis, 2019). It is often criticized for the use of the point-null hypothesis (Meehl, 1967), also known as the 'nil null' (Cohen, 1994) or 'sharp null' hypothesis (Edwards et al., 1963). It was argued that the point-null hypothesis could be appropriate only in hard sciences such as physics, but it is always false in soft sciences; this problem is sometimes known as the Meehl's paradox (Meehl, 1967, 1978; Serlin and Lapsley, 1985, 1993; Cohen, 1994; Kirk, 1996). In the case of fMRI research, we face complex brain activity which is influenced by numerous psychophysiological factors. This means that with a large amount of data, we find a statistically significant effect in all voxels for any linear contrast (Friston et al., 2002a). For example, Gonzalez-Castillo et al. (2012) showed a statistically significant difference between simple visual stimulation and rest in over 95% of the brain when averaging single-subject data from 100 runs (approximately 8 h of scanning), which consisted of five blocks of stimulation (20 s of visual stimulation, 40 s of rest). Approximately half of the brain areas showed statistically significant positive effects or 'activations,' whereas the other half showed statistically significant negative effects or 'deactivations.'

Whole-brain "activations/deactivations' can also be found when analyzing large datasets such as the HCP ($N > 1000$) or UK Biobank ($N > 10\ 000$) datasets. For example, Smith and Nichols (2018) showed significant positive and negative effects for the emotion processing task ('Emotional faces vs. Shapes' contrast) in 81% of voxels using data from UK Biobank ($N = 12\ 600$) and conservative Bonferroni multiple comparison correction. When we increase the sample size, the effect estimate does not change much. Still, the standard error in the denominator of the t-value becomes increasingly smaller, resulting in negligible effects becoming statistically significant. Thus, the classical NHST ignores the magnitude of the effect. Attempts to overcome this problem led to the proposal of making a distinction between 'statistical significance' and 'material significance' (Hodges and Lehmann, 1954) or 'practical significance' (Cohen, 1965; Kirk, 1996). That is, we can test whether the effect size is larger or smaller than some practically meaningful value using interval-null hypothesis testing (Friston et al., 2002a,b; Friston, 2013). In this case, we use the terms 'activations' and 'deactivations' for those voxels that show a practically significant positive or negative effect.

Frequentist Approach to Interval-Null Hypothesis Testing

Interval-null hypothesis testing is widely used in medicine and biology (Meyners, 2012). Consider, for example, a pharmacological study designed to compare a new treatment with an old treatment that has already shown its effectiveness. Let β_{new} be the mean effect on brain activity of the new treatment and β_{old} the mean effect of the old treatment. Then, $\theta = (\beta_{new} - \beta_{old})$ is the relative effect of the new treatment. The practical significance is defined by the effect size (ES) threshold γ. If a larger effect on brain activity is preferable, then we can test whether there is a practically meaningful difference in a positive direction (H_1: $\theta > \gamma$ vs. H_0: $\theta \leq \gamma$). This procedure is known as the *superiority test* (see **Figure 2A**). We can also test whether the effect of the new treatment is no worse (practically smaller) than the effect of the old treatment (H_1: $\theta > -\gamma$ vs. H_0: $\theta \leq -\gamma$). This procedure is sometimes known as the *non-inferiority test* (see **Figure 2B**). If a smaller effect on brain activity is preferable, we can use the superiority or non-inferiority test in the opposite direction (see **Figures 2C,D**). The combination of these two superiority tests allows us to find a practically meaningful difference in both directions (H_1: $\theta > \gamma$ and $\theta < -\gamma$ vs. H_0: $-\gamma \leq \theta \leq \gamma$), that is, the *minimum-effect test* (see **Figure 2E**). The combination of the two non-inferiority tests allows us to reject the hypothesis of practically meaningful differences in any direction (H_1: $-\gamma \leq \theta \leq \gamma$ vs. H_0: $\theta > \gamma$ and $\theta < -\gamma$). This is the most widely used approach to *equivalence testing*, known as the *two one-sided tests* (TOST) procedure (see **Figure 2F**). For more details on the superiority and minimum-effect tests, see Serlin and Lapsley (1985, 1993), Murphy and Myors (1999, 2004). For more details on the non-inferiority test and TOST procedure see Schuirmann (1987), Rogers et al. (1993), Wellek (2010), Meyners (2012), Lakens (2017).

The interval $[-\gamma; \gamma]$ defines trivially small effect sizes that we consider to be equivalent to the 'null effect' for practical purposes. This interval is also known as the 'equivalence interval' (Schuirmann, 1987) or 'region of practical equivalence (ROPE)' (Kruschke, 2011). The TOST procedure, in contrast to classical NHST, allows us to assess the 'null effects.' If we reject the null hypothesis of a practically meaningful difference, we can conclude that the effect is trivially small. The TOST procedure can also be intuitively related to frequentist interval estimates, known as confidence intervals ('confidence interval approach,' Westlake, 1972; Schuirmann, 1987). Confidence intervals reflect the uncertainty in the point estimation of the parameters defined by its standard error. The confidence level of $(1 - \alpha)$ means that among infinite 'hypothetical' replications, $(1 - \alpha)$% of the confidence intervals will contain the true effect under the null. Therefore, the TOST procedure is operationally identical to considering whether the $(1 - 2\alpha)$% confidence interval falls entirely into the ROPE, as it uses two one-sided tests with an alpha level of α.

Interval-null hypothesis testing can be used in fMRI studies not only to compare the effects of different treatments. For example, we can apply superiority tests in the positive and negative directions to detect 'activated' and 'deactivated' voxels and additionally apply the TOST procedure to detect 'not activated' voxels. However, even though we can solve the Meehl's paradox and assess the 'null effects' by switching from point-null to interval-null hypothesis testing within the frequentist approach, this approach still has fundamental philosophical and practical difficulties which can be effectively addressed using Bayesian statistics.

Difficulties of the Frequentist Approach

The pitfalls of the frequentist approach have been actively discussed by statisticians and researchers for decades. Here, we briefly mention a few of the main problems associated with the frequency approach.

(1) NHST is a hybrid of Fisher's approach that focuses on the *p*-value (thought to be a measure of evidence against the null hypothesis), and Neyman-Pearson's approach that focuses on controlling false positives with the alpha level while maximizing true positives in long-run replications. These two approaches are argued to be incompatible and have given rise to several misinterpretations among researchers, for example, confusing the meaning of *p*-values and alpha levels (Edwards et al., 1963; Gigerenzer, 1993; Goodman, 1993; Royall, 1997;

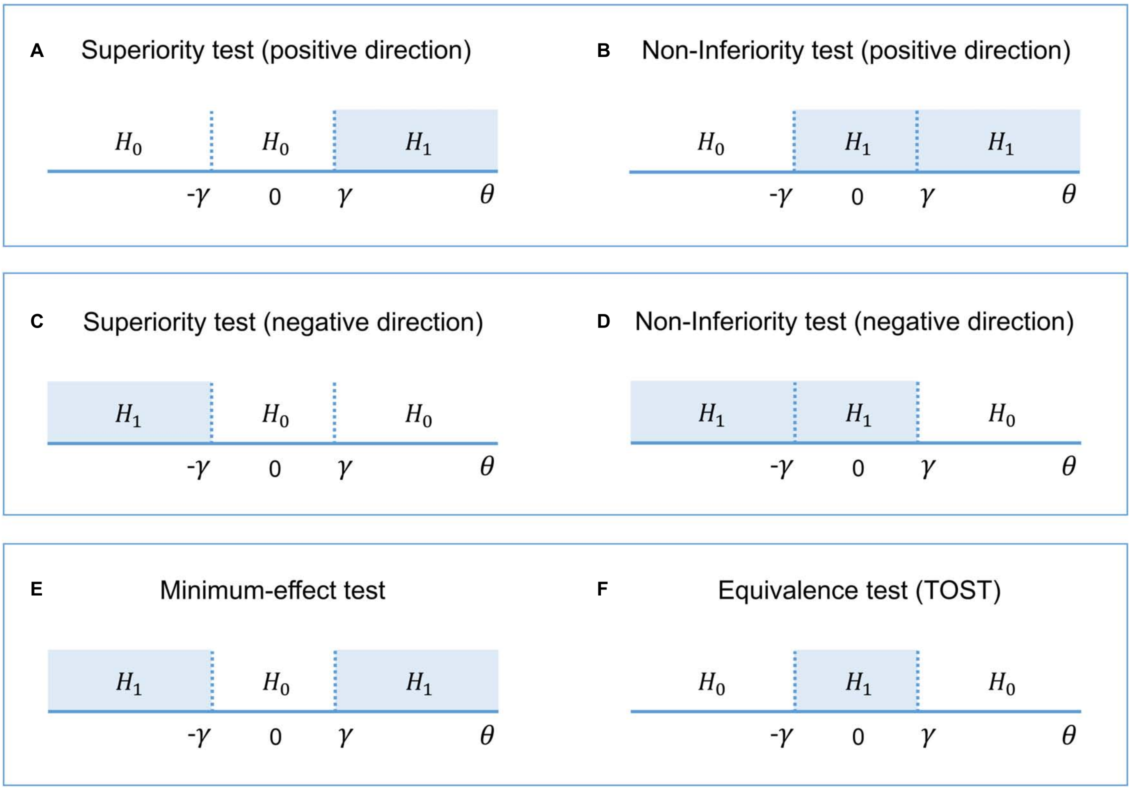

FIGURE 2 | The alternative (H_1) and null (H_0) hypotheses for different types of interval-null hypotheses tests. **(A,B)** One-sided tests in the positive direction ('the larger is better'). **(C,D)** One-sided tests in the negative direction ('the smaller is better'). **(E)** Combination of both superiority tests. **(F)** Combination of both non-inferiority tests.

Finch et al., 2001; Berger, 2003; Hubbard and Bayarri, 2003; Turkheimer et al., 2004; Schneider, 2014; Perezgonzalez, 2015; Szucs and Ioannidis, 2017; Greenland, 2019).

(2) The logical structure of NHST is the same as that of 'proof by contradiction' or 'indirect proof,' which becomes formally invalid when applied to probabilistic statements (Pollard and Richardson, 1987; Cohen, 1994; Falk and Greenbaum, 1995; Nickerson, 2000; Sober, 2008; Schneider, 2014, 2018; Wagenmakers et al., 2017; but see Hagen, 1997). Valid 'proof by contradiction' can be expressed in syllogistic form as: (1) 'If A, then B' (Premise No 1), (2) 'Not B' (Premise No 2), (3) 'Therefore not A' (Conclusion). Probabilistic 'proof by contradiction' in relation to NHST can be formulated as: (1) 'If H_0 is true, then $D+$ are highly unlikely, (2) '$D+$ was obtained,' (3) 'Therefore H_0 is highly unlikely.' This problem is also referred to as the 'illusion of probabilistic proof by contradiction' (Falk and Greenbaum, 1995). To illustrate the fallacy of such logic, consider the following example from Pollard and Richardson (1987): (1) 'If a person is an American (H_0), then he is most probably not a member of Congress,' (2) 'The person is a member of Congress,' (3) 'Therefore the person is most probably not an American.' Based on this, one 'rejects the null' and makes an obviously wrong inference, as only American citizens can be a member of Congress. At the same time, using Bayesian statistics, we can show that the null hypothesis ('the person is an American') is true (see the Bayesian solution of the 'Congress example' in the **Supplementary Materials**). The 'illusion of probabilistic proof by contradiction' leads to widespread confusion between the probability of obtaining the data, or more extreme data, under the null $P(D+|H_0)$ and the probability of the null under the data $P(H_0|\mathrm{D})$ (Pollard and Richardson, 1987; Gigerenzer, 1993; Cohen, 1994; Falk and Greenbaum, 1995; Nickerson, 2000; Finch et al., 2001; Hoekstra et al., 2006; Goodman, 2008; Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2017). The latter is a posterior probability calculated based on Bayes' rule. The fact that researchers usually treat the *p*-value as a continuous measure of evidence (the Fisherian interpretation) only exacerbates this problem. 'The lower the *p*-value, the stronger the evidence against the null' statement can be erroneously transformed to statements such as 'the lower the *p*-value, the stronger the evidence for the alternative' or 'the higher the *p*-value, the stronger the evidence for the null.' NHST can only provide evidence *against*, but never *for*, a hypothesis. In contrast, posterior probability provides direct evidence for a hypothesis; hence, it has a simple intuitive interpretation.

(3) The *p*-value is not a plausible measure of evidence (Berger and Berry, 1988; Berger and Sellke, 1987; Cornfield, 1966; Goodman, 1993; Hubbard and Lindsay, 2008; Johansson, 2011; Royall, 1986; Wagenmakers, 2007; Wagenmakers et al., 2008, 2017; Wasserstein and Lazar, 2016;

bet see Greenland, 2019). The frequentist approach considers infinite 'hypothetical' replications of the experiment (sampling distribution); that is, the *p*-value depends on unobserved ('more extreme') data. One of the most prominent theorists of Bayesian statistics, Harold Jeffreys, put it as follows: '*What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*' (Jeffreys, 1939/1948, p. 357). In turn, the sampling distribution depends on the researcher's intentions. These intentions may include different kinds of *multiplicities*, such as multiple comparisons, double-sided comparisons, secondary analyses, subgroup analyses, exploratory analyses, preliminary analyses, and interim analyses of sequentially obtained data with optional stopping (Gopalan and Berry, 1998). Two researchers with different intentions may obtain different *p*-values based on the same dataset. The problem is that these intentions are usually unknown. When null findings are considered disappointing, it is tempting to increase the sample size until one obtains a statistically significant result. However, a statistically significant result may arise when the null is, in fact, true, which can be shown by Bayesian statistics. That is, the *p*-value usually exaggerates evidence against the null hypothesis. The discrepancy that may arise between frequentist and Bayesian inference is also known as the Jeffreys–Lindley paradox (Jeffreys, 1939/1948; Lindley, 1957). In addition, it is argued that a consistent measure of evidence should not depend on the sample size (Cornfield, 1966). However, identical *p*-values provide different evidence against the null hypothesis for small and large sample sizes (Wagenmakers, 2007). In contrast, evidence provided by posterior probabilities and Bayes factors depends only on the exact observed data and the prior, and does not depend on the testing or stopping intentions or the sample size (Wagenmakers, 2007; Kruschke and Liddell, 2017b).

(4) Although frequentist interval estimates (Cohen, 1990, 1994; Cumming, 2013) and interval-based hypothesis testing (Murphy and Myors, 2004; Wellek, 2010; Meyners, 2012; Lakens, 2017) greatly facilitate the mitigation of the abovementioned pitfalls in data interpretation, they are still subject to some of the same types of problems as the *p*-values and classic NHST (Cortina and Dunlap, 1997; Nickerson, 2000; Belia et al., 2005; Wagenmakers et al., 2008; Hoekstra et al., 2014; Morey et al., 2015; Greenland et al., 2016; Kruschke and Liddell, 2017a). Confidence intervals also depend on unobserved data and the intentions of the researcher. Moreover, the meaning of confidence intervals seems counterintuitive to many researchers. For example, one of the most common misinterpretations of the $(1 - \alpha)\%$ confidence interval is that the probability of finding an effect within the confidence interval is $(1 - \alpha)\%$. In fact, it is a Bayesian interval estimate known as a *credible* interval.

Nevertheless, we would like to emphasize that we do not advocate abandoning the frequency approach. Correctly interpreted frequentist interval-based hypothesis testing with *a priori* power analysis defining the sample size and proper multiplicity adjustments often lead to conclusions similar to those of Bayesian inference (Lakens et al., 2018). However, it may be logically and practically difficult to carry out an appropriate power analysis and make multiplicity adjustments (Berry and Hochberg, 1999; Cramer et al., 2015; Streiner, 2015; Schönbrodt et al., 2017; Sjölander and Vansteelandt, 2019). These procedures may be even more complicated in fMRI research than in psychological or social studies (see discussion on power analysis in Mumford and Nichols, 2008; Joyce and Hayasaka, 2012; Mumford, 2012; Cremers et al., 2017; Poldrack et al., 2017; multiple comparisons in Nichols and Hayasaka, 2003; Nichols, 2012; Eklund et al., 2016; and other types of multiplicities in Turkheimer et al., 2004; Chen et al., 2018, 2019, 2020; Alberton et al., 2020). For example, at the beginning of a long-term study, one may want to check whether stimulus onset timings are precisely synchronized with fMRI data collection and perform preliminary analysis on the first five subjects. The question of whether the researcher must make an adjustment for this technical check when reporting the results for the final sample become important in the frequentist approach. Such preliminary analyses (or other forms of interim analyses) are generally not considered a source of concern in Bayesian inference because posterior probabilities do not depend on the sampling plan (for discussion, see Berry, 1988; Berger and Berry, 1988; Edwards et al., 1963; Wagenmakers, 2007; Kruschke and Liddell, 2017b; Rouder, 2014; Schönbrodt et al., 2017). Or, for example, one may want to find both 'activated/deactivated' and 'not activated' brain areas and use two superiority tests in combination with the TOST procedure. It is not trivial to make appropriate multiplicity adjustments in this case. In contrast, Bayesian inference suggests a single decision rule without the need for additional adjustments. Moreover, to our knowledge, practical implementations of superiority tests and the TOST procedure in common software for fMRI data analysis do not yet exist. At the same time, Bayesian analysis has already been implemented in SPM12[2] and is easily accessible to end-users. It consists of two steps: Bayesian parameter estimation and Bayesian inference. In general, it is not necessary to use Bayesian analysis at the subject level of analysis to apply it at the group level. One can combine computationally less demanding frequentist parameter estimation for single subjects with Bayesian estimation and inference at the group level. In the next sections, we consider the group-level Bayesian analysis implemented in SPM12.

Bayesian Parameter Estimation

Bayesian statistics is based on Bayes' rule:

$$P(H\,|\,D) = \frac{P(D\,|\,H)\,P(H)}{P(D)} \tag{2}$$

where $P(H|D)$ is the probability of the hypothesis given the obtained data or posterior probability. $P(D|H)$ is the probability of obtaining the *exact* data given the hypothesis or the likelihood (notice the difference from $P(D+|H)$, which includes *more extreme* data). $P(H)$ is the prior probability of the hypothesis (our knowledge of the hypothesis before we obtain the data). $P(D)$ is a normalizing constant ensuring that the sum of posterior probabilities over all possible hypotheses equals one (marginal likelihood). In the case of mutually exclusive hypotheses, the denominator of Bayes's rule is the

[2] https://www.fil.ion.ucl.ac.uk/spm/software/spm12

sum of the probabilities of obtaining the data under any of the possible hypotheses, multiplied by its prior probability. For example, if we consider two mutually exclusive hypotheses H_0 and H_1, then $P(D) = P(D|H_0)\ P(H_0) + P(D|H_1)P(H_1)$ and $P(H_0|D) + P(H_1|D) = 1$. When we consider continuous hypotheses, the denominator is obtained by integrating over all hypotheses (parameter spaces). For relatively simple models, these integrals can be solved analytically. However, for more complex models, the integrals become analytically intractable. In this case, there are two main approaches to obtain the posterior probability: (1) use computationally demanding numerical integration (Markov chain Monte Carlo methods); (2) use less accurate but computationally efficient analytical approximations to the posterior distribution (e.g., Expectation Maximization or Variational Bayes techniques). Describing these procedures go beyond the scope of this paper and described elsewhere (for their implementations in fMRI analysis, see Genovese, 2000; Friston et al., 2002a, 2007; Friston and Penny, 2003; Penny et al., 2003, 2005, 2007; Penny and Ridgway, 2013; Woolrich et al., 2004).

In verbal form, Bayes' rule can be expressed as:

$$Posterior \propto Likelihood \times Prior$$

This means that we can update our prior beliefs about the hypothesis based on the obtained data.

One of the main difficulties in using Bayesian statistics, in addition to the computational complexity, is the choice of appropriate prior assumptions. The prior can be chosen based on theoretical arguments or from independent experimental data (full Bayes approach). At the same time, if the data are organized hierarchically, which is the case for neuroimaging data, priors can be specified based on the obtained data itself using an empirical Bayes approach. The lower level of the hierarchy corresponds to the experimental effects at any given voxel, and the higher level of the hierarchy comprises the effect over all voxels. Thus, the variance of the experimental effect over all voxels can be used as the prior variance of the effect at any given voxel. This approach is known as the parametric empirical Bayes (PEB) with the 'global shrinkage' prior (Friston and Penny, 2003). The prior variance is estimated from the data under the assumption that the prior probability density corresponds to a Gaussian distribution with zero mean. In other words, a global experimental effect is assumed to be absent. An increase in local activity can be detected in some brain areas; a decrease can be found in others, but the total change in neural metabolism in the whole brain is approximately zero. This is a reasonable physiological assumption because studies of brain energy metabolism have shown that the global metabolism is 'remarkably constant despite widely varying mental and motoric activity' (Raichle and Gusnard, 2002), and 'the changes in the global measurements of blood flow and metabolism' are 'too small to be measured' by functional imaging techniques such as PET and fMRI (Gusnard and Raichle, 2001).

Now, we can rewrite Bayes' rule (eq. 2) for the effect θ = cβ:

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)} \tag{3}$$

In the process of Bayesian updating with the 'global shrinkage' prior, the effect estimate 'shrinks' toward zero. The greater the uncertainty of the effect estimate (variability) in a particular voxel, the less confidence in this estimate, and the more it shrinks (see **Figure 3**).

The assumption of a Gaussian prior, likelihood, and posterior essentially reduces computational demands for Bayesian analysis. However, the normality assumption can be violated for empirical data. For example, violations can be observed in the presence of outliers, particularly with small sample sizes or unbalanced designs, which diminishes the validity of the statistical analysis. This problem is not specific to Bayesian analysis but is inherent to all group-level analyses that assume a normal distribution of the effect. Nevertheless, in fMRI studies, the most common approach is to use the Gaussian general linear models (Poline and Brett, 2012), which have been shown to be robust against violations of the normality assumption (Knief and Forstmeier, 2021). Still, we need to be ensured that these assumptions are not violated substantially. If that is the case, one can use Bayesian estimation based on non-Gaussian distributions. In this work, we consider Bayesian estimation with Gaussian 'global shrinkage' prior implemented in SPM12.

After Bayesian parameter estimation, we can apply one of the two main types of Bayesian inference (Penny and Ridgway, 2013): *Bayesian parameter inference (BPI)* or *Bayesian model inference (BMI)*. BPI is also known as Bayesian parameter estimation (Kruschke and Liddell, 2017b). However, we deliberately separate these two terms, as they correspond to two different steps of data analysis in SPM12. BMI is also known as Bayesian model comparison, Bayesian model selection, or Bayesian hypothesis testing (Kruschke and Liddell, 2017b). We chose the term BMI as it is consonant with the term BPI.

Bayesian Parameter Inference

The BPI is based on the posterior probability of finding the effect within or outside the ROPE. Let effects larger than the ES threshold γ be 'activations,' those smaller than $-\gamma$ be 'deactivations,' and those falling within the ROPE $[-\gamma; \gamma]$ be 'no activations.' Then, we can classify voxels as 'activated,' 'deactivated,' or 'not activated' if:

$$P_{act} = P(\theta > \gamma \mid D) \geq P_{thr} \tag{4.1}$$

$$P_{deact} = P(\theta < -\gamma \mid D) \geq P_{thr} \tag{4.2}$$

$$P_{null} = P(-\gamma \leq \theta \leq \gamma \mid D) \geq P_{thr} \tag{4.3}$$

where P_{thr} is the posterior probability threshold (usually P_{thr} = 95%). Note that $P_{act} + P_{deact} + P_{null} = 1$.

If none of the above criteria are satisfied, the data in a particular voxel are insufficient to distinguish voxels that are 'activated/deactivated' from those that are 'not activated.' Hereinafter, we refer to them as 'low confidence' voxels (Magerkurth et al., 2015). This decision rule is also known as the 'ROPE-only' rule (Kruschke and Liddell, 2017a), see also Greenwald (1975); Wellek (2010), Liao et al. (2019). To the

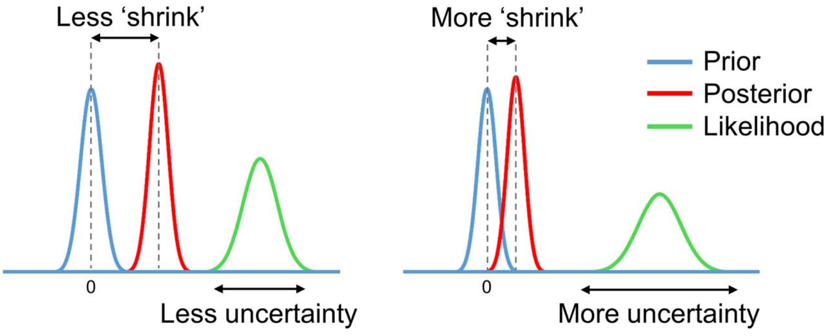

FIGURE 3 | Schematic of Bayesian updating with the 'global shrinkage' prior.

best of our knowledge, the application of this decision rule to neuroimaging data was pioneered by Friston et al. (2002a; 2002b; Friston and Penny, 2003). For convenience and visualization purposes, we can use the natural logarithm of the posterior probability odds (LPO), for example:

$$LPO_{null} = ln\left(\frac{P_{null}}{P_{act} + P_{deact}}\right) = ln\left(\frac{P_{null}}{1 - P_{null}}\right) \qquad (5)$$

This allows us to more effectively discriminate voxels with a posterior probability close to unity (Penny and Ridgway, 2013). $LPO_{null} > 3$ corresponds to $P_{null} > 95\%$. In addition, *LPO* also allows us to identify the connection between BPI and BMI. The maps of the *LPO* are termed posterior probability maps (PPMs) in SPM12.

Another possible decision rule considers the overlap between ROPE and the 95% highest density interval (HDI). HDI is a type of credible interval (Bayesian analog of the confidence interval), which contains only the effects with the highest posterior probability density. If the HDI falls entirely inside the ROPE, we can classify voxels as 'not activated.' In contrast, if the HDI lies completely outside the ROPE, we can classify voxels as either 'activated' or 'deactivated.' If the HDI overlaps with the ROPE, we cannot make a confident decision (we can consider them to be 'low confidence' voxels). This decision rule is known as the 'HDI+ROPE' rule (Kruschke and Liddell, 2017a). It is more conservative than the 'ROPE-only' rule because it does not consider the effects from the low-density tails of the posterior probability distribution. Differences between the 'HDI+ROPE' rule and the 'ROPE-only' are most evident for strongly skewed distributions. In such cases, the ROPE may contain more than 95% of the posterior probability distribution, but the 95% HDI may overlap with the ROPE. In the case of a Gaussian posterior probability distribution, both decision rules should produce similar results. The 'HDI+ROPE rule is advocated by Kruschke and Liddell (2017a) and the 'ROPE-only' rule is preferred by Friston et al. (2002a; 2002b; Friston and Penny, 2003), Wellek (2010); Liao et al. (2019). These decision rules are illustrated in **Figure 4**.

Bayesian Model Inference

With BPI, we consider the posterior probabilities of the linear contrast of parameters $\theta = c\beta$. Instead, we can consider models using BMI.

Let H_{alt} and H_{null} be two non-overlapping hypotheses represented by models M_{alt} and M_{null}. These models are defined by two parameter spaces: (1) M_{alt}: $\theta > \gamma$ and $\theta < -\gamma$, and (2) M_{null}: $-\gamma \leq \theta \leq \gamma$.

Now, we can rewrite Bayes' rule (eq. 2) for M_{alt} and M_{null}

$$P\left(M_{alt} \mid D\right) = \frac{P\left(D \mid M_{alt}\right) P\left(M_{alt}\right)}{P\left(D\right)} \qquad (6.1)$$

$$P\left(M_{null} \mid D\right) = \frac{P\left(D \mid M_{null}\right) P\left(M_{null}\right)}{P\left(D\right)} \qquad (6.2)$$

If we divide equation (6.1) by (6.2), $P(D)$ is canceled out, and we obtain:

$$\frac{P\left(M_{alt} \mid D\right)}{P\left(M_{null} \mid D\right)} = \frac{P\left(D \mid M_{alt}\right)}{P\left(D \mid M_{null}\right)} \frac{P\left(M_{alt}\right)}{P\left(M_{null}\right)} \qquad (7)$$

In verbal form equation (7) can be expressed as:

Posterior Odds = Bayes Factor × Prior Odds

The Bayes factor (*BF*) is a multiplier that converts prior model probability odds to posterior model probability odds. It indicates the relative evidence for one model against another. For example, if $BF_{null} = \frac{p(D|M_{null})}{p(D|M_{alt})} = 2$, then the observed data are twice as likely under the null model than under the alternative.

A connection exists between the BPI (eq. 3–5), and BMI (eq. 7) (see Morey and Rouder, 2011; Liao et al., 2019):

$$BF_{null} = \left(\frac{P(-\gamma \leq \theta \leq \gamma|D)}{1 - P(-\gamma \leq \theta \leq \gamma|D)}\right)\left(\frac{1 - P(-\gamma \leq \theta \leq \gamma)}{P(-\gamma \leq \theta \leq \gamma)}\right) \qquad (8)$$

or, in verbal form:

$$BF(ROPE)_{null} = \frac{Posterior(\theta \in ROPE)}{Posterior\ (\theta \notin ROPE)} \frac{Prior(\theta \notin ROPE)}{Prior(\theta \in ROPE)}$$

For convenience, *BF* may also be expressed in the form of a natural logarithm:

$$LogBF(ROPE)_{null} = LPO_{null} + ln\left(\frac{Prior(\theta \notin ROPE)}{Prior(\theta \in ROPE)}\right) \qquad (9)$$

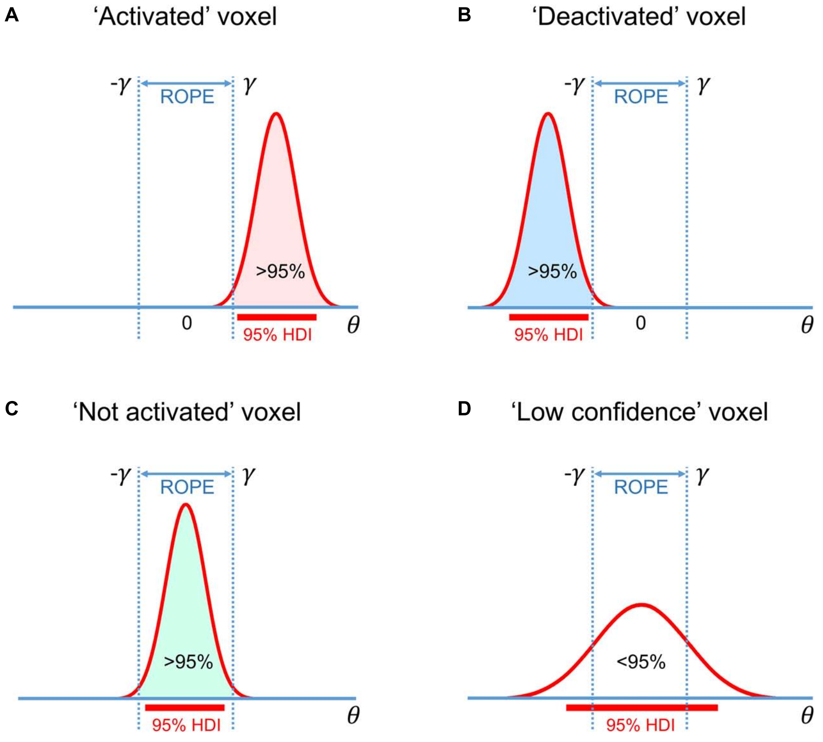

FIGURE 4 | Possible variants of the posterior probability distributions of the effect θ = *c*β in **(A)** 'activated' voxels, **(B)** 'deactivated' voxels, **(C)** 'not activated' voxels, **(D)** 'low confidence' voxels. The 'ROPE only' rule considers only the colored parts of the distributions. The 'HDI+ROPE' rule considers overlap between the ROPE and 95% HDI.

$$logBF(ROPE)_{null} \propto LPO_{null} \tag{10}$$

The calculation of *BF* may be computationally challenging, as it requires integration over the parameter space. However, if the ROPE has zero width (point-null hypothesis), then the *BF* has an analytical solution known as the Savage–Dickey ratio (SDR) (Wagenmakers et al., 2010; Friston and Penny, 2011; Rosa et al., 2012; Penny and Ridgway, 2013). $BF(SDR)_{null}$ is calculated by dividing the prior probability density by the posterior probability density at θ = 0. The interpretation of the $BF(SDR)_{null}$ is simple: if the effect size is less likely to equal zero after obtaining the data than before, then $BF(SDR)_{null} < 1$: that is, we have more evidence for M_{alt}. See schematic illustration of BMI based on interval-null and point-null hypotheses and its relation to BPI in **Figure 5**.

Relations Between Frequentist and Bayesian Approaches

Now we can point out the conceptual links between the frequentist and Bayesian approaches.

(1) **Parameter estimation**: When we have no prior information, that is, all parameter values are *a priori* equally probable ('flat' prior), the PEB estimation reduces to the frequentist parameter estimation (maximum likelihood estimation; Friston et al., 2002a).

(2) **Multiplicity adjustments**: One of the major concerns in frequentist inference is the multiplicity problem. In general, after the Bayesian parameter estimation, it is not necessary to classify any voxel as 'activated/deactivated ' or 'not activated.' If we consider *unthresholded* maps of posterior probabilities, *LPOs*, or *LogBFs*, the multiple comparisons problem does not arise (Friston and Penny, 2003). However, if we apply a decision rule to classify voxels, we should control for wrong decisions across multiple comparisons (Woolrich et al., 2009, see also possible loss functions in Muller et al., 2006; Kruschke and Liddell, 2017a). The advantage of PEB with the 'global shrinkage' prior is that it automatically accounts for multiple comparisons without the need for *ad hoc* multiplicity adjustments (Berry, 1988; Friston and Penny, 2003; Gelman et al., 2012). The frequentist approach processes every voxel independently, whereas the PEB algorithm considers joint information from all voxels. Frequentist inference uncorrected for multiple independent comparisons is prone to label noise-driven, random extremes as 'statistically significant.' Bayesian analysis specifies that extreme values are unlikely *a priori*, and thus they shrink toward a common mean (Lindley, 1990; Westfall et al., 1997; Berry and Hochberg, 1999; Friston et al., 2002a,b; Gelman et al., 2012; Kruschke and Liddell, 2017b). If we consider *thresholded* maps of posterior probabilities, for example, $P_{act} > 95\%$, then as many as 5% of 'activated' voxels could be falsely labeled so. This is conceptually similar to the false discovery rate (FDR) correction (Berry and Hochberg, 1999; Friston et al., 2002b; Friston and Penny, 2003; Storey, 2003; Muller et al., 2006; Schwartzman et al., 2009). In practice, BPI with γ = 0 should produce similar results (in terms of the number

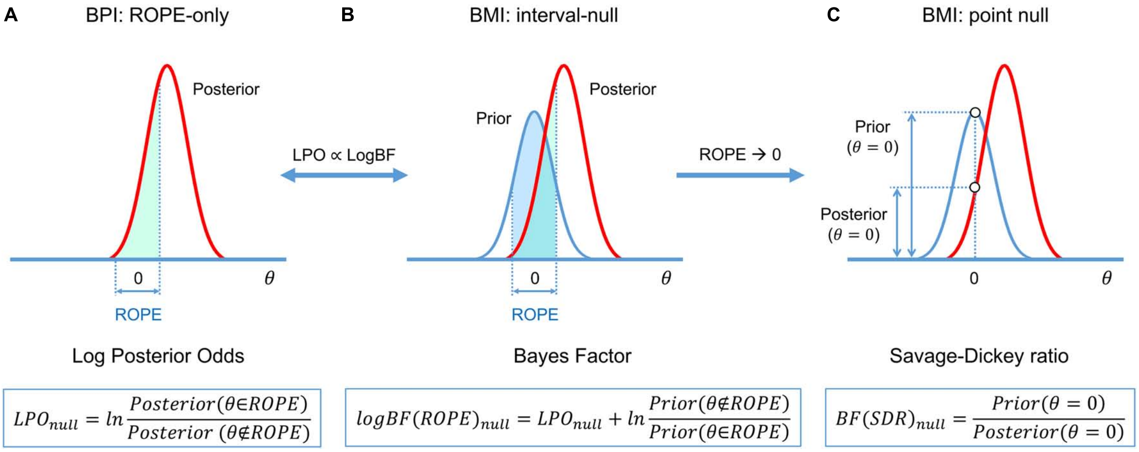
A
BPI: ROPE-only
B
BMI: interval-null
C
BMI: point null
Posterior
Prior
Posterior
Prior
$(\theta = 0)$
Posterior
$(\theta = 0)$
LPO ∝ LogBF
ROPE → 0
0
θ
ROPE
0
θ
ROPE
0
θ
Log Posterior Odds
Bayes Factor
Savage-Dickey ratio
$LPO_{null} = ln\frac{Posterior(\theta \in ROPE)}{Posterior\ (\theta \notin ROPE)}$
$logBF(ROPE)_{null} = LPO_{null} + ln\frac{Prior(\theta \notin ROPE)}{Prior(\theta \in ROPE)}$
$BF(SDR)_{null} = \frac{Prior(\theta = 0)}{Posterior(\theta = 0)}$

FIGURE 5 | Schematic of *BFs* used in BMI and their relation to *LPO* used in BPI. **(A)** BPI based on the 'ROPE-only' decision rule. **(B)** *BF(ROPE)* is related to the areas under the functions of the posterior and prior probability densities inside and outside the ROPE. **(C)** *BF(SDR)* is the relation between the posterior and prior probability at θ = 0. *LPOs* and *BFs* provide relative evidence for the null and alternative hypotheses.

of 'activated/deactivated' voxels) as classical NHST with FDR correction. If we increase the ES threshold, fewer voxels will be classified as 'activated/deactivated,' and at some γ value, BPI will produce results similar to the more conservative Family Wise Error (FWE) correction[3].

(3) **Interval-based hypothesis testing**: Frequentist interval-based hypothesis testing is conceptually connected with BPI, particularly, the 'HDI+ROPE' decision rule. The former considers the intersection between ROPE and the confidence intervals. The latter considers the intersection between ROPE and the HDI (credible intervals).

(4) **BPI and BMI**: BMI based on *BF(ROPE)* is conceptually linked to BPI based on the 'ROPE-only' decision rule. The interval-based Bayes factor *BF(ROPE)* is proportional to the posterior probability odds. When ROPE is infinitesimally narrow, *BF* can be approximated using the *SDR*. Note that even though *BF(SDR)* is based on the point-null hypothesis, it can still provide evidence for the null hypothesis, in contrast to BPI with $\gamma = 0$. However, *BF(SDR)* in PEB settings has not yet been tested using empirical fMRI data. Because the point-null hypothesis is always false (Meehl, 1967), BPI and *BF(ROPE)* may be preferred over *BF(SDR)*.

Definition of the Effect Size Threshold

The main difficulty in applying interval-based methods is the choice of the ES threshold γ. To date, only a few studies have been devoted to determining the minimal relevant effect size. One of them suggested a method to objectively define γ at the subject level of analysis which was calibrated by clinical experts and may be implemented for pre-surgical planning (Magerkurth et al., 2015). At the same time, the problem of choosing the ES threshold γ for the group-level Bayesian analysis remains unresolved.

Several ways in which to define the ES threshold are available. Firstly, we can conduct a pilot study to determine the expected effect sizes. Secondly, we can use data from the literature to determine the typical effect sizes for the condition of interest. Thirdly, we can use the default ES thresholds that are commonly accepted in the field. One of the first ES thresholds proposed in the neuroimaging literature was $\gamma = 0.1\%$ (Friston et al., 2002b). This is the default ES threshold for the subject-level BPI in SPM12. For the group-level BPI, the default ES threshold is one prior standard deviation of the effect $\gamma = 1\ prior\ SD_\theta$ (Friston and Penny, 2003). Fourthly, γ can be selected in such a way as to ensure maximum similarity of the activation patterns revealed by classical NHST and Bayesian inference. This would allow us to reanalyze the data using Bayesian inference, reveal similar activation patterns as was previously the case for classic inference, and detect the 'not activated' and 'low confidence' voxels. Lastly, we can consider the posterior probabilities at multiple ES thresholds or compute the ROPE maps (see below).

The ES threshold can be expressed as unstandardized (raw β values or percent signal change) and standardized values (for example, Cohen's d). Raw β values calculated by SPM12 at the first level of analysis represent the BOLD signal in arbitrary units. However, they can be scaled to a more meaningful unit, the BOLD percent signal change (PSC) (Poldrack et al., 2011; Chen et al., 2017). Unstandardized and standardized values have disadvantages and advantages. Different ways exist in which to scale β values to PSC (Pernet, 2014; Chen et al., 2017), which is problematic when comparing the results of different studies. Standardized values represent the effect size in terms of the standard deviation units, which supposedly facilitate the comparison of results between different experiments. However, standardized values are relatively more unstable between measurements and less interpretable (Baguley, 2009; Chen et al., 2017). Moreover, Cohen's d is closely related to the t-value (for one sample case, $d = \ t/\sqrt{N}$) and may share some drawbacks with t-values. Reimold et al. (2005) showed that spatial smoothing

[3] FDR correction controls the rate of false discoveries (false positives in frequentist terminology) among all significant voxels. FWE correction controls the rate of any false positives in the whole brain.

has a nonlinear effect on voxel variance. Using t-values or Cohen's d for inference in neuroimaging may lead to spatially inaccurate results (spatial shift of local maxima in t-maps or Cohen's d maps compared to PSC-maps). In this study, we focused on PSCs.

It is also important to note that effect sizes (both BOLD PSC and Cohen's d) depend on the MRI scanner (e.g., field strength, coil sensitivity), acquisition parameters (e.g., echo time, spin echo vs. gradient echo sequences) and field inhomogeneity (UIudag et al., 2009). For example, the effect sizes may be underestimated in brain areas near air–tissue interfaces because of field inhomogeneities. This fact further complicates the selection of the ES threshold. However, this does not mean that we should ignore the effect size and return to the point-null hypothesis. One may choose different ES thresholds for different regions of interest, scanners or acquisition parameters.

METHODS

Datasets

Seven block-design tasks were considered from the HCP dataset, including working memory, gambling, motor, language, social cognition, relation processing, and emotion processing tasks (Barch et al., 2013). Two event-related tasks, including the stop-signal and task-switching tasks were considered from the UCLA dataset (Poldrack et al., 2016). The length, conditions, and number of scans of the tasks are provided in the **Supplementary Materials** (**Supplementary Table 1**). A subset of 100 unrelated subjects (S1200 release) was selected from the HCP dataset (54 females, 46 males, mean age = 29.1 ± 3.7 years) for assessment. A total of 115 subjects from the UCLA dataset were included in the analysis (55 females, 60 males, mean age = 31.7 ± 8.9 years) after removing subjects with no data for the stop-signal task, a high level (>15%) of errors in the Go-trials, and those of which the raw data were reported to be problematic (Gorgolewski et al., 2017). See the fMRI acquisition parameters in the **Supplementary Materials**, Par. 1.

Preprocessing

The minimal preprocessing pipelines for the HCP and UCLA datasets were described by Glasser et al. (2013) and Gorgolewski et al. (2017), respectively. Spatial smoothing was applied to the preprocessed images with a 4 mm full width at half maximum (FWHM) Gaussian smoothing kernel. Additionally, to compare the extent to which the performance of classical NHST and BPI depended on the smoothing, we applied 8 mm FWHM smoothing to the emotion processing task. Spatial smoothing was performed using SPM12. The results are reported for the 4 mm FWHM smoothing filter, unless otherwise specified.

Parameter Estimation

Frequentist parameter estimation was applied at the subject level of analysis. A detailed description of the general linear models for each task design is available in the **Supplementary Materials**, Par. 2. Fixation blocks and null events were not modeled explicitly in any of the tasks. Twenty-four head motion regressors were included in each subject-level model (six head motion parameters, six head motion parameters one time point before, and 12 corresponding squared items) to minimize head motion artifacts (Friston et al., 1996). Raw β values were converted to PSC relative to the mean whole-brain 'baseline' signal (**Supplementary Materials**, Par. 3). The linear contrasts of the β values were calculated to describe the effects of interest $\theta = c\beta$ in different tasks. The sum of positive terms in the contrast vector, c, is equal to one. The list of contrasts calculated in the current study to explore typical effect sizes is presented in **Supplementary Table 1**. At the group level of analysis, the Bayesian parameter estimation with the 'global shrinkage' prior was applied using SPM12 (v6906). We performed a one-sample test on the linear contrasts created at the subject level of analysis.

Classical Null Hypothesis Significance Testing and Bayesian Parameter Inference

Classical inference was performed using voxel-wise FWE correction with $\alpha = 0.05$. This is the default SPM threshold and is known to be conservative and to guarantee protection from false positives (Eklund et al., 2016). Although voxel-wise FWE correction may be too conservative for small sample sizes, it is recommended when large sample sizes are available (Woo et al., 2014).

Bayesian parameter inference, accessible via the SPM12 GUI, allows the user to declare only whether the voxels are 'activated' or 'deactivated.' The classification of voxels as being either 'not activated' or 'low confidence' requires the posterior mean and variance. At the group level of analysis, SPM12 does not save the posterior variance image. However, the posterior variance can be reconstructed from the image of the noise hyperparameter using a first-order Taylor series approximation (Penny and Ridgway, 2013). Therefore, in the current study, BPI was performed using the developed SPM12-based toolbox[4]. For the 'ROPE-only' rule, the posterior probability threshold was $P_{thr} = 95\%$ ($LPO > 3$). For the 'HDI+ROPE' rule, we used the 95% HDI.

We compared the number of 'activated' voxels (as a percentage of the total number of voxels) detected by Bayesian and classical inference. We also compared the number of 'activated,' 'deactivated,' and 'not activated' voxels detected using BPI with the 'ROPE-only' and 'HDI+ROPE' decision rules and different ES thresholds. To estimate the influence of the sample size on the results, all the above-mentioned analyses were performed with samples of different sizes: 5 to 100 subjects from the HCP dataset (the emotion processing task, 'Emotion > Shape' contrast) and 5 to 115 subjects from the UCLA dataset (the stop signal task, 'Correct Stop > Go' contrast), in steps of 5 subjects. Ten random groups were sampled for each step.

Effect Size Thresholds

We considered three ES thresholds: firstly, the default ES threshold for the subject-level $\gamma = 0.1\%$ (BOLD PSC); secondly, the default ES threshold for the group-level $\gamma = 1$ *prior* SD_θ; thirdly, the $\gamma(Dice_{max})$ threshold, which ensures maximum

[4] https://github.com/Masharipov/Bayesian_inference

similarity of the activation patterns revealed by classical NHST and BPI. The similarity was assessed using the Dice coefficient:

$$Dice\,(\gamma) = \frac{2 * V_{overlap}\,(\gamma)}{V_{classic} + V_{bayesian}\,(\gamma)} \qquad (11)$$

where $V_{classic}$ is the number of 'activated' voxels detected using classical NHST, $V_{bayesian}\,(\gamma)$ is the number of 'activated' voxels detected using BMI with the ES threshold γ, and $V_{overlap}$ is the number of 'activated' voxels detected by both methods. A Dice coefficient of 0 indicates no overlap between the patterns, and 1 indicates complete overlap. Dice coefficients were calculated for γ ranging from 0 to 0.4% in steps of 0.001%.

Estimation of Typical Effect Sizes

In the current study, we aimed to provide a reference set of typical effect sizes for different task designs (block and event-related) and different contrasts ('task-condition > control-condition,' 'task-condition > baseline,') in a set of *a priori* defined regions of interest (ROI). Effect sizes were expressed in PSC and Cohen's d. ROI masks were defined using anatomical and *a priori* functional masks. For more details, see **Supplementary Materials**, Par. 4.

Evaluating Bayesian Parameter Inference on Contrasts With No Expected Practically Significant Difference

Bayesian parameter inference should be able to detect the 'null effect' in the majority of voxels when comparing samples with no expected practically significant difference. For example, there may be two groups of healthy adult subjects performing the same task or two sessions with the same task instructions. To test this, we used fMRI data from the emotion processing task. To emulate two 'similar' *independent* samples, 100 healthy adult subjects' contrasts ('Emotion > Shape') were randomly divided into two groups of 50 subjects. A two-sample test comparing the 'Group #1' and 'Group #2' was performed with the assumption of unequal variances between the groups (SPM12 default option). To emulate 'similar' *dependent* samples, we randomized 'Emotion > Shape' contrasts from right-to-left (RL) and left-to-right (LR) phase encoding sessions in the 'Session #1' and 'Session #2' samples. Each sample consisted of 50 contrasts from the RL session and 50 from the LR session. A paired test designed to compare 'Session #1' and 'Session #2' was equivalent to the one-sample test on 50 'RL > LR session' and 50 'LR > RL session' contrasts.

Normality Check

To check for violations of the normality assumption we performed Shapiro-Wilk test (Shapiro and Wilk, 1965) for each voxel for one block-design task ('Emotion >Shape' contrast) and one event-related task ('Correct Stop > Go' contrast). We reported the number of voxels that were significantly non-Gaussian (using $\alpha = 0.001$ uncorrected for multiple comparisons and $\alpha = 0.05$ with Bonferroni correction). We also calculated median kurtosis and skewness across voxels. Kurtosis is a measure of the heaviness of the tails. Skewness is a measure of asymmetry of distribution.

Simulations

The main limitation of using empirical data to assess the performance of statistical methods lies in the lack of knowledge of the ground truth. Therefore, we performed group-level simulations to better understand how the sample size and effect size threshold affect BPI results given different known effect sizes and noises. Simulations also allowed us to assess the robustness of BPI to the violations of the normality assumption. We generated the parameter maps (contrast images) similar to Nichols and Hayasaka (2003); Schwartzman et al. (2009) and Cremers et al. (2017). Each contrast image consisted of 'activated' and 'deactivated' voxels and 'trivial' background voxels surrounding them. Locations of 'activated' and 'deactivated' voxels were specified based on the NeuroSynth meta-analysis results (Yarkoni et al., 2011) obtained using the search terms 'task' and 'default,' respectfully (association test, $\alpha = 0.01$ with FDR correction). Data were drawn from the Pearson system distribution (Johnson et al., 1994) with kurtosis, $Ku = 2.2, 3, 7$ and skewness, $Sk = -0.7, 0, 0.7$. The normal distribution corresponds to $Ku = 3$ and $Sk = 0$. Other combinations of Ku and Sk resulted in four-parameter beta distributions. The mean effect in practically significant ('activated' and 'deactivated') voxels was $\theta = \pm\,0.1, 0.2, 0.3\%$. For practically non-significant or 'trivial' voxels, the mean effect was $\theta = \pm\,0.04\%$, which can be considered equivalent to the null value for practical purposes ('not activated' voxels). Noise standard deviation was $SD = 0.2, 0.3, 0.4\%$. The mean effect size and noise were consistent with those observed in the empirical data (see **Supplementary Tables 11–19**). Contrast-to-noise ratio was varied from 0.25 to 1.5. For each combination of the Pearson system distribution parameters, we generated 1000 images.

To evaluate sample size dependencies, we randomly drawn images from the full sample ($N = 1000$) ranging from $N = 10$ to 100 (with step 10) and from $N = 150$ to 500 (step 50). This procedure was repeated ten times for each step. The analysis was limited to the single axial slice ($z = 36$ mm) containing 579 'activated' voxels, 500 'deactivated' voxels and 3067 'trivial' or 'not activated' voxels. For classical NHST and BPI, we calculated the number of 'activated' voxels in relation to the total number of voxels. For BPI, we additionally calculated:

(1) Correct decision rate. The number of correctly classified 'activated,' 'deactivated,' and 'not activated' voxels to its true number (c.f. 'hit rate' in detection theory or 'true positive rate' in frequentist framework).
(2) Incorrect decision rate. The number of voxels incorrectly classified as 'activated,' 'deactivated,' and 'not activated' to the true number of voxels not belonging to 'activated,' 'deactivated,' and 'not activated' categories, respectfully (c.f. 'false alarm rate' in detection theory or 'false positive rate' in frequentist framework);
(3) Low confidence decision rate. The number of 'low confidence' voxels to the total number of voxels.

The code for the simulations is available online[5].

[5] https://github.com/Masharipov/BPI_2021/tree/main/simulations

RESULTS

Results for Contrasts With No Expected Practically Significant Difference

Classical NHST did not show a significant difference between 'Group #1' and 'Group #2' (see **Supplementary Figure 1**). BPI with the 'ROPE-only' decision rule and default ES threshold $\gamma = 1$ *prior* $SD_\theta = 0.190\%$ classified 83.4% of voxels as having 'no difference' in which the null hypothesis was accepted (see **Supplementary Figure 1**). The 'HDI+ROPE' rule classified 76.2% of voxels as having 'no difference.'

Classical NHST did not reveal a significant difference between 'Session #1' and 'Session #2' (see **Supplementary Figure 2**). The *prior* SD_θ was 0.005%. In this case, using the default ES threshold $\gamma = 1$ *prior* SD_θ did not allow the detection of any 'no difference' voxels, because the ROPE was unreasonably narrow. The 'null effect' was detected in all voxels beginning with a $\gamma = 0.013\%$ threshold using the 'ROPE-only' and 'HDI+ROPE' decision rules (see **Supplementary Figure 2**).

In this way, when comparing two 'similar' *independent* samples (two groups of healthy subjects performing the same task), BPI with the default group-level threshold (*one prior* SD_θ) allowed us to correctly label voxels as having 'no difference' for the majority of the voxels of the brain. However, when comparing two 'similar' *dependent* samples (two sessions from the same task), the *one prior* SD_θ threshold became inadequately small.

Therefore, the default *one prior* SD_θ threshold is not suitable when the difference between *dependent* conditions is very small (paired sample test or one-sample test). In such cases, one can use an *a priori* defined ES threshold based on previously reported effect sizes or provide an ES threshold at which most of the voxels can be labeled as having 'no difference,' allowing the critical reader to decide whether this speaks in favor of the absence of differences.

Comparison of Classical Null Hypothesis Significance Testing and Bayesian Parameter Inference Results

Generally, classical NHST with voxel-wise FWE correction and BPI with the 'ROPE-only' decision rule and default group-level ES threshold $\gamma = 1$ *prior* SD_θ revealed similar (de)activation patterns in all considered contrasts (see **Figure 6**, **Table 1**, and **Supplementary Tables 2–10**). The median ES threshold based on $Dice_{max}$ and median default group-level ES threshold across all considered contrasts were close in magnitude to the default subject-level ES threshold $\gamma = 0.1\%$: $\gamma(Dice_{max}) = 0.118\%$ and $\gamma = 1$ *prior* $SD_\theta = 0.142\%$. The median $Dice_{max}$ across all the considered contrasts reached 0.904. At the same time, BPI allowed us to classify 'non-significant' voxels as 'not activated' or 'low confidence.' As it can be clearly seen from **Figure 6**, areas with 'non-activated' voxels surround clusters with 'activated/deactivated' voxels. Both are separated by areas comprising 'low confidence' voxels.

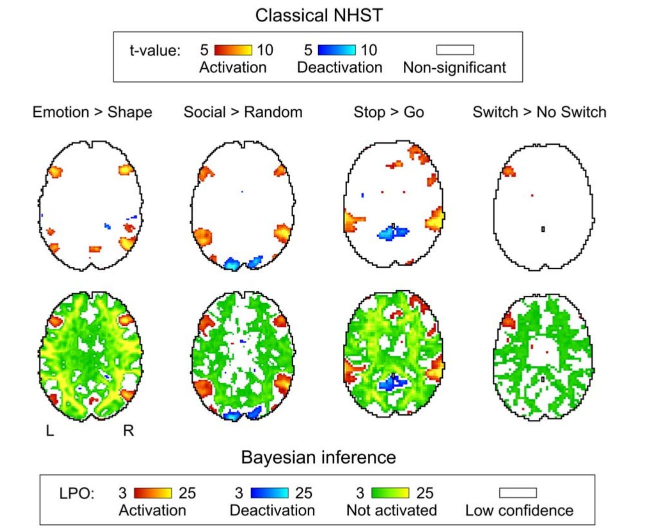

FIGURE 6 | Examples of results obtained with classical NHST and BPI. Four contrasts were chosen for the illustration purposes (two event-related and two block-design tasks). Classical NHST was implemented using voxel-wise FWE correction (α = 0.05). BPI was implemented using the 'ROPE-only' decision rule, $P_{thr} = 95\%$ ($LPO > 3$) and $\gamma = 1$ *prior* SD_θ. Axial slice $z = 18$ mm (MNI152 standard space).

As expected, compared with the 'HDI+ROPE' rule, using the 'ROPE-only' rule slightly increases the number of 'activated/deactivated' and 'not activated' voxels (see **Table 1** and **Supplementary Tables 2–10**). The 'HDI+ROPE' rule labeled more voxels as 'low confidence.'

Comparison of Bayesian Parameter Inference Results With Different Effect Size Thresholds

Here, we focus on the 'ROPE-only' rule. We first consider the results for the emotional processing task and then consider other tasks. Using the default single-subject ES threshold $\gamma = 0.1\%$ for the emotional processing task ('Emotion > Shape' contrast), 58.8% of all voxels can be classified as 'not activated,' 30.8% as 'low confidence,' and 10.1% as 'activated' (see **Figure 7** and **Supplementary Table 2**). The default group-level ES threshold $\gamma = 1$ *prior* $SD_\theta = 0.135\%$ allowed us to classify 75.0% of all voxels as 'non-activated,' 17.5% as 'low confidence,' and 7.4% as 'activated' (see **Figure 7** and **Supplementary Table 2**). Both types of thresholds were comparable to those of classical NHST for the detection of 'activated' voxels. The maximum overlap between 'activations' patterns revealed by classical NHST and BPI was observed at $\gamma(Dice_{max}) = 0.116\%$ (see **Figure 8** and **Table 1**).

For the '2-back > 0-back,' 'Left Finger > baseline,' 'Right Finger > baseline,' and 'Social > Random' contrasts, the three ES thresholds that were considered—0.1%, *one prior* SD_θ, $\gamma(Dice_{max})$—produced similar results (see **Table 1** and **Supplementary Tables 3**, **5**, **7**). For the event-related stop-signal task ('Correct Stop > baseline' and 'Correct Stop > Go' contrasts), *one prior* SD_θ and $\gamma(Dice_{max})$ were close in terms of their values but smaller than 0.1% (see **Table 1**). Block designs tend to evoke higher BOLD PSC than event-related designs; therefore,

TABLE 1 | Maximum Dice coefficient and corresponding effect size thresholds for each task.

Contrast, θ	*Prior* SD_θ, %	'ROPE-only' decision rule		'HDI+ROPE' decision rule	
		$\gamma(Dice_{max})$, %	$Dice_{max}$	$\gamma(Dice_{max})$, %	Dice
Emotion processing					
Emotion > Shape	0.135	0.116	0.904	0.104	0.912
Working memory					
2-back > baseline	0.325	0.136	0.925	0.125	0.932
2-back > 0-back	0.089	0.095	0.891	0.089	0.903
Language					
Story > Math	0.255	0.119	0.896	0.108	0.904
Motor					
Left finger > baseline	0.149	0.148	0.897	0.135	0.907
Right finger > baseline	0.171	0.160	0.886	0.144	0.897
Tongue > baseline	0.268	0.205	0.904	0.181	0.913
Gambling					
Reward > baseline	0.254	0.132	0.917	0.122	0.924
Loss > baseline	0.249	0.134	0.918	0.118	0.925
Reward > Loss	0.032	0.044	0.894	0.037	0.886
Social cognition					
Social > baseline	0.325	0.139	0.939	0.124	0.944
Social > Random	0.104	0.114	0.896	0.104	0.907
Relational processing					
Relational > baseline	0.390	0.154	0.935	0.143	0.940
Relational > Match	0.051	0.073	0.892	0.066	0.894
Stop-signal task					
Correct Stop > baseline	0.069	0.066	0.895	0.061	0.906
Correct Stop > Go	0.064	0.052	0.906	0.047	0.917
Task-switching					
Switch > baseline	0.133	0.075	0.907	0.067	0.916
Switch > No switch	0.030	0.037	0.924	0.033	0.925
Summary					
Median	0.142	0.118	0.904	0.106	0.913

a lower *prior* SD_θ should be expected for event-related designs and higher *prior* SD_θ for block designs. Within a single design, in contrasts such as 'task-condition > baseline,' higher BOLD PSC and *prior* SD_θ would be expected than in contrasts in which the experimental conditions are compared directly. For example, the contrast '2-back > baseline' has *prior* $SD_\theta = 0.325\%$ and contrast '2-back > 0-back' has *prior* $SD_\theta = 0.089\%$.

As previously noted, some contrasts did not elicit robust activations: 'Reward > Loss,' 'Relational > Match,' (Barch et al., 2013) and 'Switch > No switch' (Gorgolewski et al., 2017). The corresponding $\gamma(Dice_{max})$ thresholds were 0.044, 0.073, and 0.037% (see **Table 1** and **Supplementary Tables 6**, **8**, **10**). The *prior* SD_θ were 0.032, 0.051, and 0.030%. Correspondingly, BPI with the $\gamma = 1$ *prior* SD_θ threshold classified 0, 18.4, and 42.2% of voxels as 'not activated.' This demonstrates that when we compare conditions with similar neural activity and minor differences, it becomes more difficult to separate 'activations/deactivations' from the 'null effects' using the $\gamma = 1$ *prior* SD_θ threshold.

Typical Effect Sizes in Functional Magnetic Resonance Imaging Studies

A complete list of effect sizes (BOLD PSC and Cohen's d) estimated for different tasks and *a priori* defined ROIs is presented in the **Supplementary Materials** (**Supplementary Tables 11–19**). Here, we focus only on the BOLD PSC. The violin plots for some of these are shown in **Figure 9**.

For example, the median BOLD PSC in the left amygdala ROI, one of the key brain areas for emotional processing, was 0.263%, which is approximately twice as large as *one prior* SD_θ (see **Figure 7**). Thus, using this PSC as the ES threshold in future studies may cause the ROPE to become too wide compared to the effect sizes typical for tasks with such designs. Therefore, such a threshold can be used to detect large and highly localized effects. However, it may fail to detect small but widely distributed effects previously described for HCP data (Cremers et al., 2017).

In general, median PSCs within ROIs were up to 1% for block designs and 0.5% for event-related designs. The maximum PSCs reached 2.5% and were usually observed in the primary visual cortex (V1) for visual tasks comparing

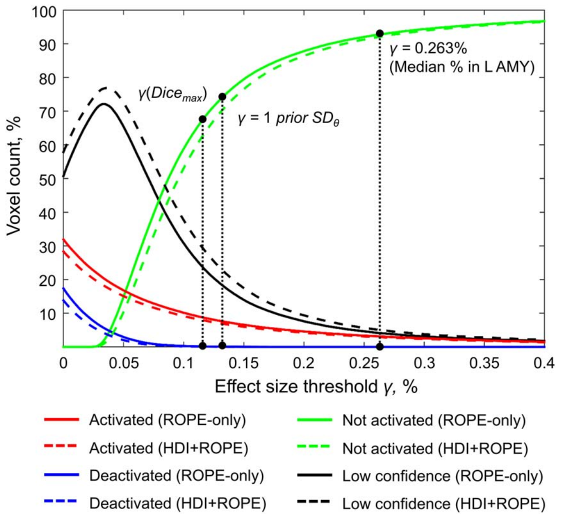

FIGURE 7 | Number of voxels classified into the four categories depending on the ES threshold γ. The results for the emotion processing task ('Emotion>Shape' contrast) are presented for illustration. L AMY, left amygdala.

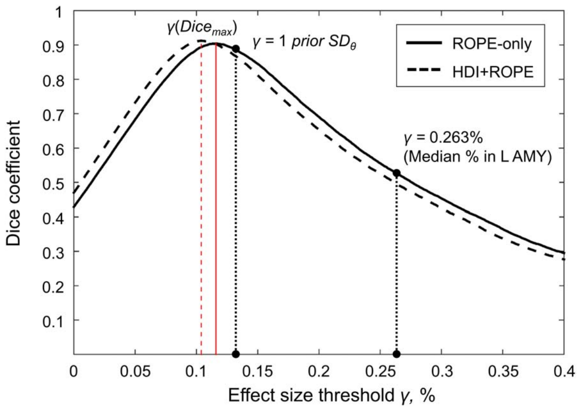

FIGURE 8 | Dependence of the Dice coefficient on the ES threshold γ. Results for the emotion processing task ('Emotion>Shape' contrast). The red lines denote γ(*Dice*$_{max}$). L AMY, left amygdala.

experimental conditions with baseline activity. For 'moderate' physiological effects, PSC varied in the range 0.1—0.2%, for example, for the '2-back > 0-back' contrast, the median PSC in the right dorsolateral prefrontal cortex (R DLPFC in **Figure 9**) was 0.137%. Likewise, for the 'Social > Random' contrast, the right inferior parietal lobule (R IPL) median PSC was 0.137%, for the 'Correct Stop > Go,' the right inferior frontal gyrus/frontal operculum (R IFG/FO) median

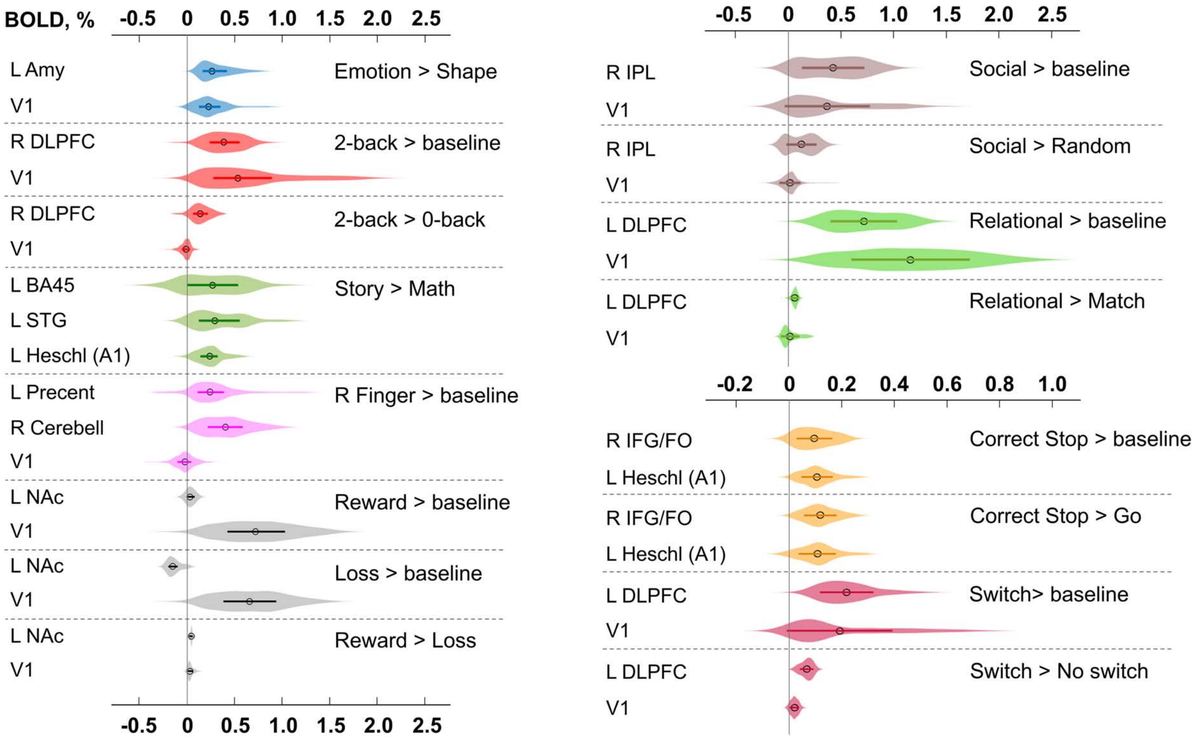

FIGURE 9 | Typical BOLD PSC in fMRI studies. The box plots inside the violins represent the first and third quartile, and the black circles represent median values. Contrasts from the same task are indicated in one color. L/R, left/right; AMY, amygdala; V1, primary visual cortex; DLPFC, dorsolateral prefrontal cortex; BA, Brodmann area; STG, superior temporal gyrus; A1, primary auditory cortex; NAc, nucleus accumbens; IPL, inferior parietal lobule; IFG/FO, inferior frontal gyrus/frontal operculum.

PSC was 0.120%. For more 'strong' physiological effects, the PSC was in the range 0.2−0.3%, for example, for the 'Emotion > Shape' contrast, the median PSC in the left amygdala was 0.263%, and for the 'Story > Math' contrast, the median PSC in the left Brodmann area 45 (Broca's area) was 0.269%. For the motor activity, for example the 'Right Finger > baseline' contrast, the median PSC in the left precentral gyrus was 0.239%, in the left postcentral gyrus was 0.362%, in the left putamen was 0.290%, and in the right cerebellum was 0.401%. For the contrasts that did not elicit robust activations (Barch et al., 2013), the PSC was approximately 0.05–0.1%; for example, for the 'Reward > Loss' contrast, the median PSC in the left nucleus accumbens was 0.043%, and for the 'Relational > Match' contrast, the median PSC in the left dorsolateral prefrontal cortex was 0.062%.

Region of Practical Equivalence Maps

We considered BPI with two consecutive thresholding steps: (1) calculate the *LPOs* (or PPMs) with a selected ES threshold γ, (2) apply the posterior probability threshold p_{th} = 95% or consider the overlap between the 95% HDI and ROPE. We can *reverse the thresholding sequence* and calculate *the ROPE maps*.

For the 'activated/deactivated' voxels, the ROPE map contains the maximum ES thresholds that allow voxels to be classified as 'activated/deactivated' based on the 'ROPE-only' or 'HDI+ROPE' decision rules. For the 'not activated' voxels, the map contains the minimum effect size thresholds that allow voxels to be classified as 'not activated.'

The procedure for calculating the ROPE map can be performed as follows. Let us consider a gradual increase in the ROPE radius (i.e., the half-width of ROPE or the ES threshold γ) from zero to the maximum effect size in observed volume. (1) For voxels in which PSC is close to zero, at a certain ROPE radius, the posterior probability of finding the effect within the ROPE becomes higher than 95%. This width is indicated on the ROPE map for 'not activated' voxels. (2) For voxels in which the PSC deviates from zero, at a certain ROPE radius, the posterior probability of finding the effect outside the ROPE becomes lower than 95%. This width is indicated on the ROPE map for 'activated/deactivated' voxels. The same maps can be calculated for the 'HDI+ROPE' decision rule.

Examples of the ROPE maps are shown in **Figure 10**. From our point of view, ROPE maps, as well as unstandardized effect size (PSC) maps, may facilitate an intuitive understanding of the spatial distribution of a physiological effect under investigation (Chen et al., 2017). They can also be a valuable addition to standard PPMs, allowing researchers to flexibly choose the ES threshold based on expected effect size for specific experimental conditions, ROIs and MR acquisition parameters. The default ES thresholds may be more conservative to brain areas near air–tissue interfaces due to signal dropout. The researcher may choose a lower ES threshold to increase sensitivity to these brain areas.

Effects of Spatial Smoothing on Classical Null Hypothesis Significance Testing and Bayesian Parameter Inference

Two main effects of spatial smoothing were identified. Firstly, higher spatial smoothing increased the number of both 'activated/deactivated' and 'not activated' voxels classified by BPI, reducing the number of 'low confidence' voxels. Secondly, higher smoothing blurred the spatial localisation of local maxima of t-maps and PPMs (*LPO*-maps) to a different extent. Consider, for example, the emotion processing task ('Emotion > Shape' contrast). The broadening of two peaks in the left and right amygdala was more noticeable on the t-map than on the PPM (see **Figure 11**).

Smoothing was previously shown to have a nonlinear effect on the voxel variances and thus to affect more t-maps than β value maps, sometimes leading to counterintuitive artifacts (Reimold et al., 2005). This is especially noticeable at the border between two different tissues or between the two narrow peaks of the local maxima. If the peak is localized close to white matter voxels with low variability, then smoothing can shift the peak to the white matter. If low-variance white matter voxels separate two close peaks, then after smoothing, they may serve as a 'bridge' between the two peaks. To avoid this problem, Reimold et al. (2005) recommended using masked β value maps. In the present study, we suggest that PPMs based on BOLD PSC thresholding can mitigate this problem. Importantly, smoothing artifacts can also arise on Cohen's d maps. Therefore, PPMs based on PSC thresholding may be preferable to PPMs based on Cohen's d thresholding.

Sample Size Dependencies for Classical Null Hypothesis Significance Testing and Bayesian Parameter Inference

An enlargement of the sample size led to an increase in the number of 'activated' and 'not activated' voxels, and a decrease in the number of 'low confidence' voxels. This is due to a decrease in the posterior variance. The curve of the 'activated' voxels rose much slower than that of the 'not activated' voxels. For the emotion processing task ('Emotion > Shape' contrast, block-design, two sessions, 352 scans), the largest gain in the number of 'activated' and 'not activated' voxels can be noted from 20 to 30 subjects (see **Figure 12A**). With a sample size of $N > 30$, the number of 'activated' and 'not activated' voxels increased less steeply. The 'not activated' and 'low confidence' voxels curves intersected at $N = 30$ subjects. After the intersection point, the graphs reached a plateau.

Considering only half of the emotional processing task data (one session, 176 scans), the intersection point shifted from $N = 30$ to $N = 60$ (see **Figure 12B**). For the event-related task ('Correct Stop > Go' contrast, the stop-signal task, 184 scans), all considered dependencies had the same features as for the block-design task, and the point of intersection was at $N = 60$ subjects (see **Figure 12C**). For the fixed ES threshold, the moment at which the graphs reach a plateau depends on task design, data quality and the amount of data at the subject level, that is, on

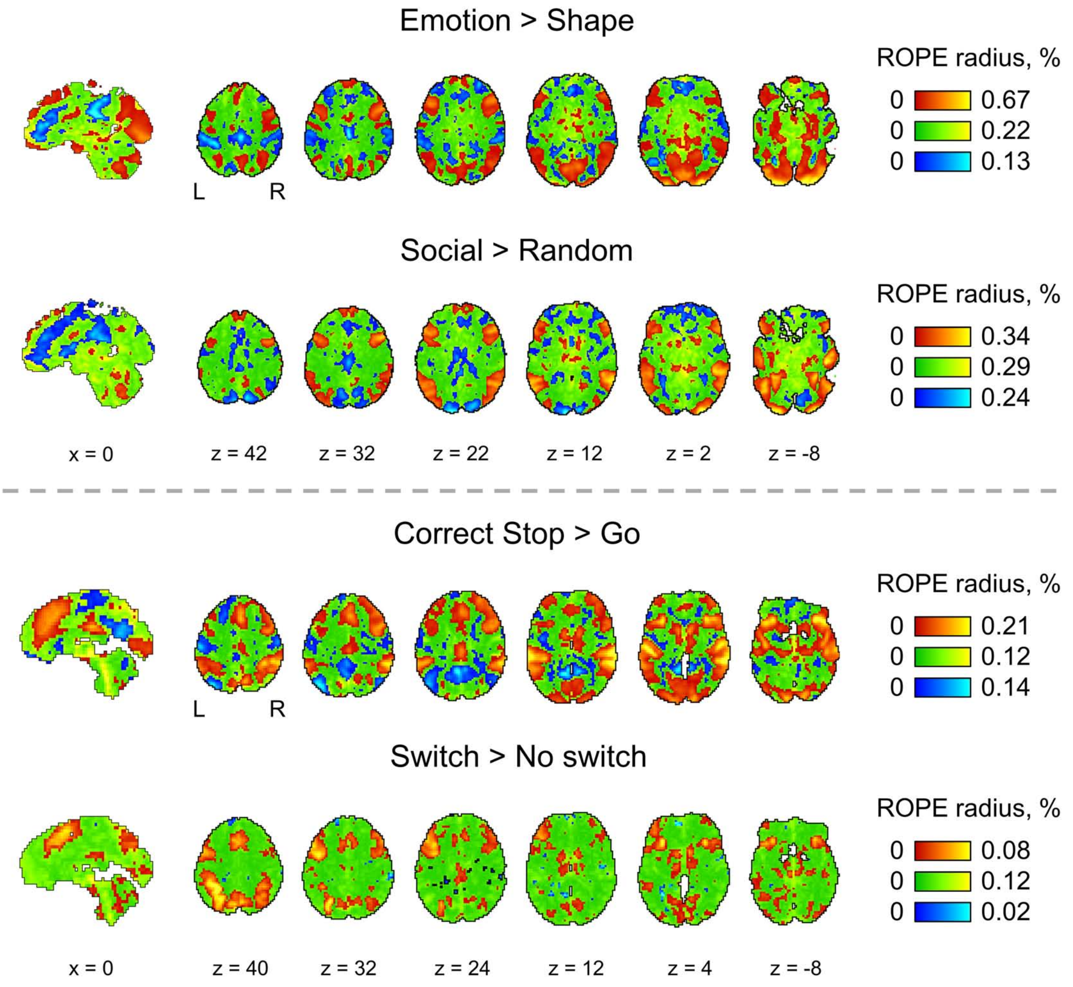

FIGURE 10 | The ROPE maps. Four contrasts were chosen for the illustration purposes (two event-related and two block-design tasks). The ROPE maps are presented using different colors for the 'activated,' 'deactivated,' and 'not activated' voxels. The green bars represent the minimum ROPE radii at which voxels with a PSC close to zero can be classified as 'not activated' based on the 'ROPE-only' decision rule. The red and blue bars represent the maximum ROPE radii at which voxels of which the PSC deviates from zero can be classified as 'activated' and 'deactivated,' respectively.

the number of scans, blocks, and events. The task designs from the HCP and UCLA datasets have relatively short durations (for example, the stop-signal task has approximately 15 'Correct Stop' trials per subject). Studies with a shorter scanning time generally require a larger sample size to enable inferences to be made with confidence. Lowering the ES threshold would also require larger sample size to reach a plateau.

Classical NHST with the voxel-wise FWE correction showed a steady linear increase in the number of 'activated' voxels with increasing sample size (see **Figure 13**). With a further increase in the sample size, the number of statistically significant voxels revealed by classical NHST is expected to approach 100% (see, for example, Gonzalez-Castillo et al., 2012; Smith and Nichols, 2018). In contrast, the BPI with the $\gamma = 1$ *prior* SD_θ threshold demonstrated hyperbolic dependencies. We observed a steeper increase at small and moderate sample sizes ($N = 15–60$). The curve of the 'activated' voxels flattened at large sample sizes ($N > 80$). BPI offers protection against the detection of 'trivial' effects that can appear as a result of an increased sample size if classical NHST with the point-null hypothesis is used (Friston et al., 2002a; Friston, 2012; Chen et al., 2017). This is achieved by the ES threshold γ, which eliminates physiologically (practically) negligible effects. **Figure 13** presents an illustration of the Jeffreys-Lindley paradox, that is, the discrepancy between results obtained using classical and Bayesian inference, which is usually manifested at higher sample sizes (Jeffreys, 1939/1948; Lindley, 1957; Friston, 2012).

Normality Check

For the block-design task ('Emotion > Shape' contrast), the number of significantly non-Gaussian voxels was 17% with $\alpha_{uncorr} = 0.001$ and 2% with $\alpha_{Bonf} = 0.05$. The median kurtosis and skewness across voxels was $Ku = 3.77$ and $Sk = 0.05$. For the event-related task ('Correct Stop > Go' contrast), the number of significantly non-Gaussian voxels was 19% with $\alpha_{uncorr} = 0.001$ and 4% with

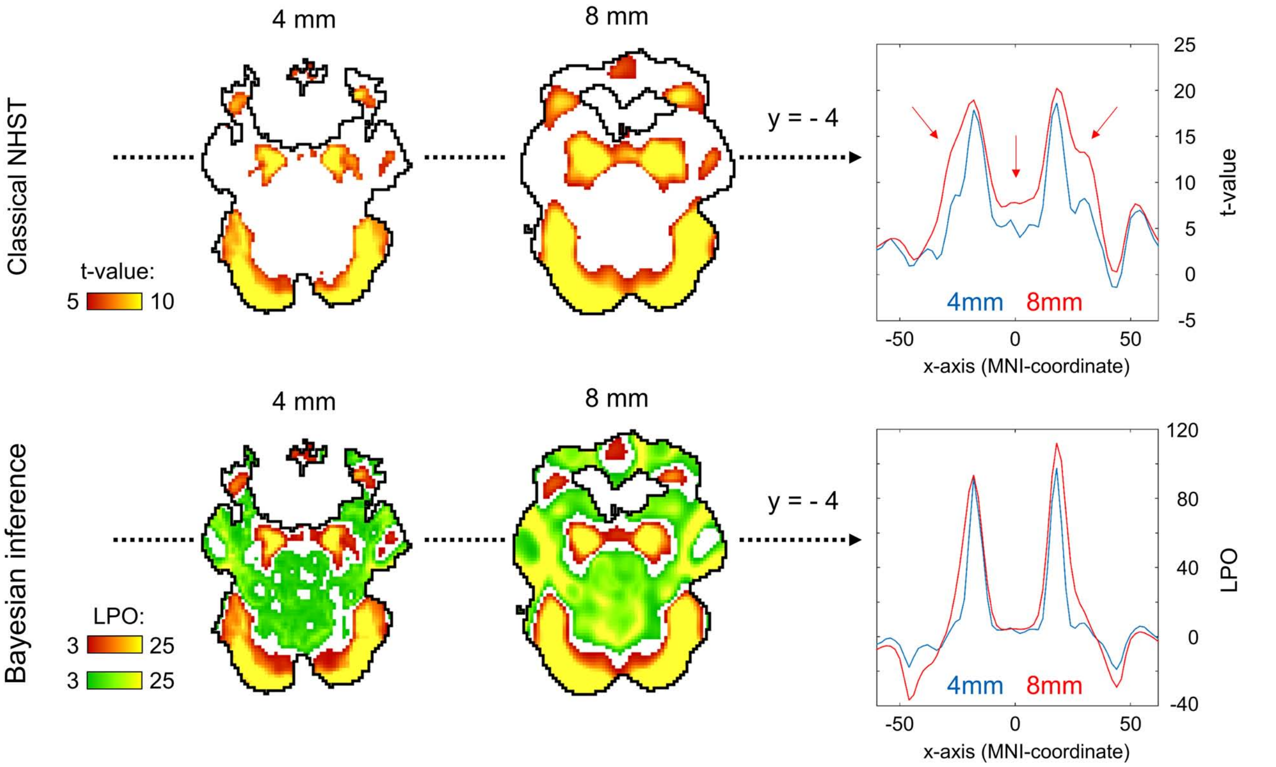

FIGURE 11 | Influence of spatial smoothing on classical NHST and BPI: results for the emotion processing task ('Emotion > Shape' contrast). Classical NHST was implemented using voxel-wise FWE correction (α = 0.05). BPI was implemented using the 'ROPE-only' decision rule, P_{thr} = 95% (*LPO* > 3) and γ = 1 *prior* SD_θ = 0.135%. Axial slice *z* = –14 mm (MNI152 standard space). Slice images have different outlines due to spatial smoothing (higher spatial smoothing increases the size of implicit masks for single subjects and group of subjects). In the panels on the right, 1-D images are presented for t-values and *LPOs* along the x-axis for *y* = –4 mm. The red arrows indicate a noticeable broadening of two peaks of local maxima (left and right amygdala) at higher smoothing.

α_{Bonf} = 0.05. The median kurtosis and skewness across voxels was *Ku* = 3.77 and *Sk* = 0.05. In general, the data are consistent with the normality assumption, though some voxels violate it.

Simulations

The simulations results reproduced the results obtained from the empirical data (see **Figure 14** for an overview of the simulations). Further, they allowed us to demonstrate how various factors affect BPI performance with the known ground truth.

Dependence of the Number of 'Activated' Voxels on the Sample Size

The number of 'activated' voxels revealed by BPI with the γ = 1 *prior* SD_θ threshold approaches the true number of practically significant voxels and stops increasing (see **Figure 15**). Classical NHST shows further increase of 'activated' voxels with the sample size increase, as it considers only statistical significance. This is more evident for low and medium noise cases (*SD* = 0.2, 0.3%). For the high noise case (*SD* = 0.4%), the sample size should be larger than *N* = 500 for the discrepancy between NHST and BPI results to become evident.

Dependence of the Correct and Low Confidence Decision Rates on the Sample Size

For the weak effect size (θ = 0.1%), the BPI with the γ = 1 *prior* SD_θ threshold is more sensitive for 'activated' than for 'not activated' voxels (see **Figure 16**). This is because γ = 1 *prior* SD_θ threshold is smaller for the weak effect size. For the moderate and strong effects (θ = 0.2, 0.3%), this difference in sensitivity become less evident. The low confidence decisions are prevalent in the 'weak effect plus high noise' case. It becomes more challenging to distinguish between 'activated' and 'not activated' voxels when the data are noisy, and the PSC in the 'activated' voxels is close to the PSC in 'trivial' voxels. For the intermediate case (moderate effect plus medium noise), the correct decision rates for 'activated' and 'not activated' voxels reached 80% at the sample sizes *N* = 80 and *N* = 150, correspondingly. For larger effect sizes and lower noise, a smaller sample size will be required to achieve the correct decision rate of 80% (and vice versa). The 'ROPE-only' decision rule is more sensitive to both 'activated' and 'not activated' voxels than the 'HDI+ROPE' decision rule.

Robustness of Bayesian Parameter Inference to Violations of the Normality Assumption

Non-normal distributions with positive and negative skewness increase incorrect decision rates for 'deactivated' and 'activated'

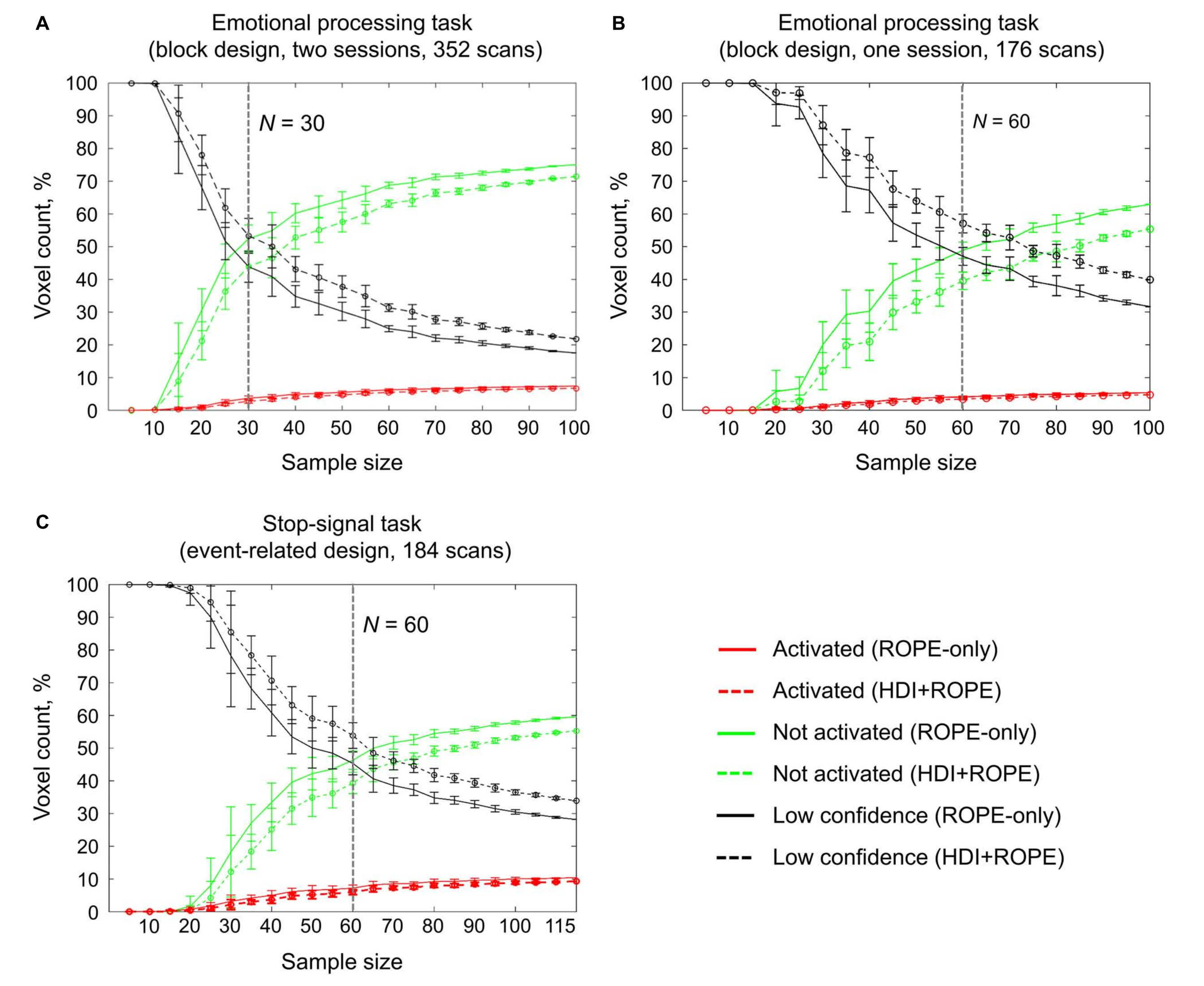

FIGURE 12 | Dependencies of the number of 'activated,' 'not activated,' and 'low confidence' voxels on the sample size. BPI was implemented using $\gamma = 1$ *prior* SD_θ. **(A)** The emotional processing task ('Emotion > Shape' contrast, two sessions). **(B)** The emotional processing task ('Emotion > Shape' contrast, one session). **(C)** The stop-signal task ('Correct Stop > Go' contrast). The error bars represent the mean and standard deviation across ten random groups.

voxels, correspondingly (**Figure 17**). Application the 'ROPE-only' decision rule results in higher incorrect decision rates than the 'HDI+ROPE' decision rule. However, even in the worst case (weak effect plus high noise), the incorrect decision rates for BPI with the $\gamma = 1$ *prior* SD_θ threshold did not exceed 5%. This result shows that BPI is robust to violations of the normality assumption. The 'ROPE-only' rule may be preferable to the 'HDI+ROPE' rule, as both rules protect against incorrect decisions, but the 'ROPE-only' rule is more sensitive to the true effects using $\gamma = 1$ *prior* SD_θ threshold.

Dependence of the Correct and Incorrect Decision Rates on the Effect Size Threshold

The optimal ES threshold should provide high sensitivity to both 'activated' and 'not activated' voxels (e.g., higher than 80%) while protecting against incorrect decisions (e.g., lower than 5%). The range of ES thresholds that meets these criteria decreases for lower true effects and higher noise (see **Figure 18**). At the sample size $N = 200$, the default $\gamma = 1$ *prior* SD_θ threshold falled in the range of optimal ES thresholds in the majority of the cases. For the weak effect plus high noise case, one should choose between high sensitivity to 'activated' or 'not activated' voxels. In this scenario, to achieve high sensitivity to both types of voxels, it is necessary to obtain a very large sample size ($N > 500$). In all considered cases, the default ES threshold provided approximately equal correct decision rates for 'activated' and 'not activated' voxels and protected against incorrect decisions. This result confirmed that the default IS threshold is optimal in most scenarios, except for the scenario with low effect and high noise level.

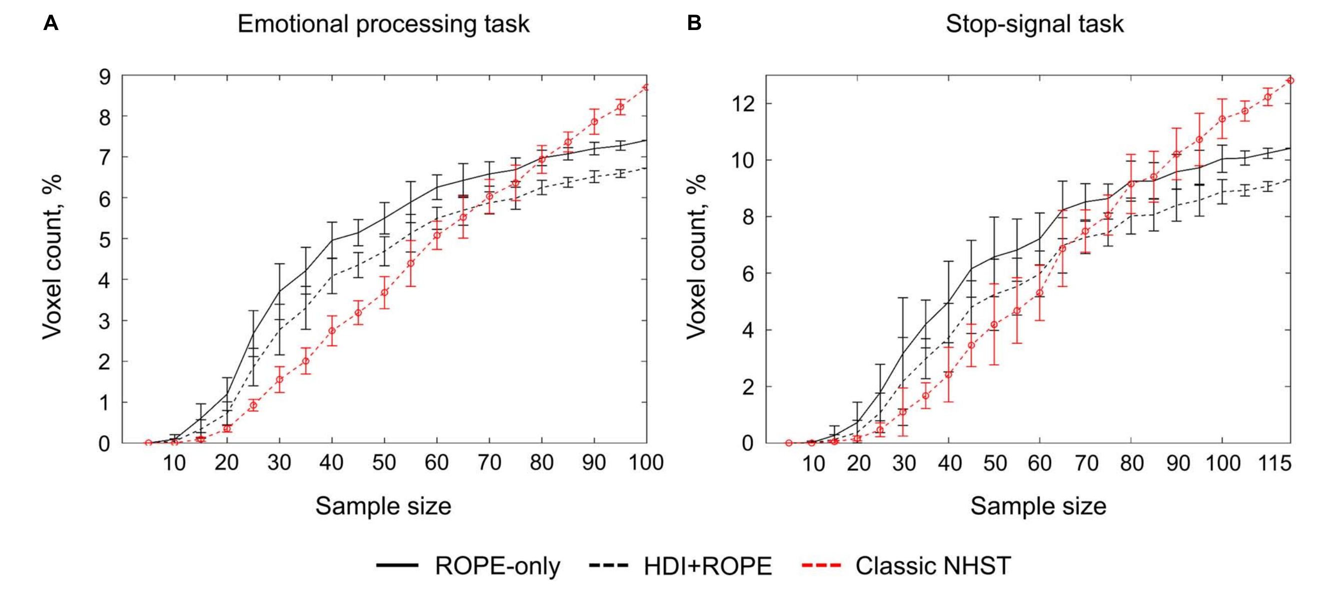

FIGURE 13 | Dependencies of the number of 'activated' voxels on the sample size. Classical NHST was implemented using FWE correction (α = 0.05). BPI was implemented using $\gamma = 1$ *prior* SD_θ. **(A)** The emotional processing task (block design, 'Emotion > Shape' contrast). **(B)** The stop-signal task (event-related design, 'Correct Stop > Go' contrast). The error bars represent the mean and standard deviation across ten random groups.

Example of Practical Application of Bayesian Parameter Inference

In contrast to classical NHST, Bayesian inference allows us to:

(1) Provide evidence that there is no practically meaningful BOLD signal change in the brain area when comparing the two task conditions.
(2) Establish double dissociations; that is to state that one area responds to A *but not* B condition and another responds to B *but not* A condition (Friston et al., 2002a).
(3) Provide evidence for practically equivalent engagement of one area under different experimental conditions in terms of local brain activity.
(4) Provide evidence for the absence of a practically meaningful difference in BOLD signals between groups of subjects or repeated measures.

To illustrate a possible application of Bayesian inference in research practice, we used a working memory task. Let us consider an overlap between the '2-back > baseline' and '0-back > baseline' contrasts (see **Figure 19**, purple areas). We cannot claim that brain areas revealed by this conjunction analysis were equally engaged in the '2-back' and '0-back' conditions. To provide evidence for this notion, we can use BPI and attempt to identify voxels with a practically equivalent BOLD signal in the '2-back' and '0-back' conditions (see **Figure 19**, green areas). Overlap between the '2-back > baseline' and '0-back > baseline' and the '2-back = 0-back' effects was found in several brain areas: visual cortex (V1, V2, V3), frontal eye field (FEF), superior eye field (SEF), parietal eye field (PEF, or posterior parietal cortex), lateral geniculate nucleus (LGN) and left primary motor cortex (M1) (see **Figure 19**, white areas). This result can be easily explained by the fact that both experimental conditions require the subject to analyze perceptually similar visual stimuli and push response buttons with the right hand, which should not depend much on the working memory load. At the same time, it does not follow directly from simple conjunction analysis.

DISCUSSION

Over-reliance on classical NHST promotes publication bias toward statistically significant findings. However, the null result can be just as valuable and exciting as the statistically significant result. Furthermore, not every statistically significant result has a practical significance. In recent years, statistical practice has seen a gradual shift from point-null hypothesis testing to interval-null hypothesis testing and interval estimation, as well as from frequentist to Bayesian approaches. Frequentist and Bayesian interval-based approaches allow us to assess the 'null effects' and thus overcome prejudice against the null hypothesis. While both approaches may lead to similar results (if specially calibrated to get it), we discussed conceptual and practical reasons for preferring the Bayesian approach. One of the main conceptual difficulties of the frequentist approach is that it is based on the probabilistic 'proof by contradiction,' which results in the 'inverse probability' fallacy: that is a widespread misinterpretation of *p*-values and confidence intervals as posterior probabilities and credible intervals. Although the Bayesian approach does not automatically guarantee correct interpretations, it can be more intuitive and straightforward than the frequentist approach (particularly, Bayesian inference based on the posterior probability distributions of the parameters or BPI).

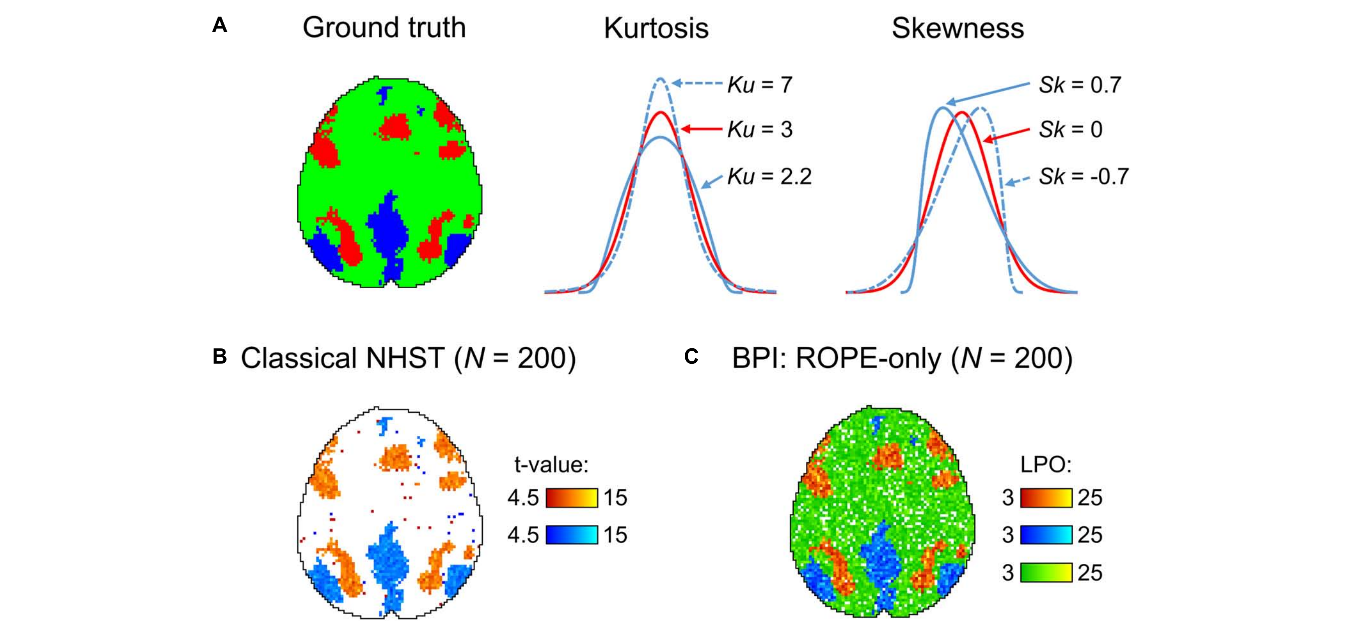

FIGURE 14 | Simulations overview. **(A)** Ground truth axial slice $z = 36$ mm (MNI152 standard space). 'Activated' and 'deactivated' voxels are marked in red and blue colors, respectfully. 'Trivial' voxels that should be classified as 'not activated' (practically equivalent to the null value) are marked in green. Data were drawn from the normal ($Ku = 3$, $Sk = 0$, the red line) and non-normal distributions. **(B)** Classical NHST results for $N = 200$ images, moderate effect and medium noise ($\theta = 0.2\%$, $SD = 0.3\%$), obtained using voxel-wise FWE correction ($\alpha = 0.05$). **(C)** BPI results for $N = 200$ images, moderate effect and medium noise ($\theta = 0.2\%$, $SD = 0.3\%$), obtained using the 'ROPE-only' decision rule, $P_{thr} = 95\%$ ($LPO > 3$) and $\gamma = 1$ *prior* SD_θ.

At the same time, from the frequentist point of view, the main conceptual disadvantage of the Bayesian approach is the need to specify our prior beliefs about the model parameters. Sometimes it is argued that we do not want our result to depend on a subjective prior decision. However, in the frequentist framework, we also make prior assumptions when subjectively choosing a model or ignoring the prior distributions of model parameters (implicitly use 'flat' prior). From this point of view, the explicit choice of the prior may be rather an advantage. We can choose prior from theoretical arguments (e.g., biophysical or anatomical priors) or derive prior from the hierarchically organized data (empirical Bayes approach). In this way, we limit the subjectivity of the choice of the prior.

Another potential obstacle to using Bayesian statistics is its computational complexity. Integrals in Bayes' rule can be solved analytically only for relatively simple models. In other cases, numerical integration approaches should be used to calculate the posterior probability, which are particularly time-consuming, especially when considering thousands of voxels. Alternatively, one can use computationally efficient analytical approximations to the posterior distributions, which, however, can be less accurate for high-dimensional parameter spaces (multivariate analysis).

Despite profound development of Bayesian techniques, to date, the 'null effect' assessment is uncommon in neuroimaging field and, in particular, in fMRI studies. One of the possible reasons for this may be the lack of tools available to the end-user. To facilitate the 'null effect' assessment for fMRI practitioners, we developed SPM12 based toolbox for group-level Bayesian inference[4]. We evaluated the BPI approach on empirical and simulated data and discussed its possible application in fMRI studies.

Bayesian parameter inference allows us to simultaneously find 'activated/deactivated,' 'not activated,' and 'low confidence' voxels using a single decision rule. The 'not activated' decision means that the effect is practically non-significant and can be considered equivalent to the null for practical purposes. The 'low confidence' decision means we need more data to make a confident inference, that is, we need to increase the scanning time, sample size, data quality or revisit the task design. The use of parametrical empirical Bayes with the 'global shrinkage' prior enables us to check the results as the sample size increases and allows us to decide whether to stop the experiment if the obtained data are sufficient to make a confident inference. All the above features are absent from the classical NHST framework, limited to the point-null hypothesis with a pre-determined stopping rule.

An important advantage of Bayesian inference is that we can use graphs such as those shown in **Figure 12** to determine when the obtained data are sufficient to make a confident inference. We can plot such graphs for the whole brain or for *a priori* defined ROIs. When the curves reach a plateau, the data collection can be stopped. If the brain area can be labeled as either 'activated/deactivated' or 'not activated' at a relatively small sample size, it will be still so at larger sample sizes. If the brain area can be labeled as 'low confidence,' we must increase the sample size to make a confident inference. At a certain sample size, it could possibly be labeled as either 'activated/deactivated' or 'not activated.' In the worst case, we can

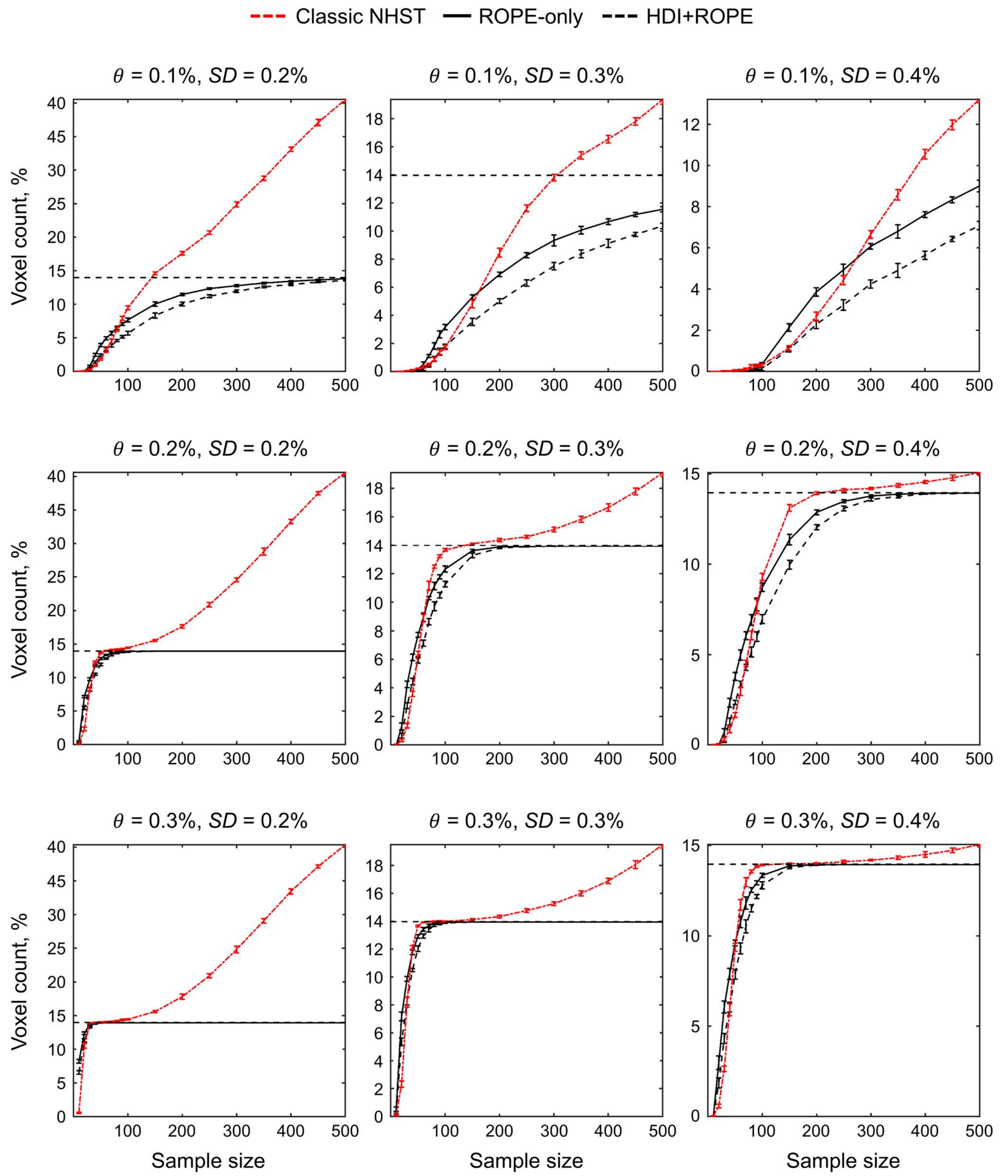

FIGURE 15 | Simulations results for the dependencies of the number of 'activated' voxels on the sample size. Data were drawn from normal distributions with different mean effect θ and noise *SD*. Classical NHST was implemented using FWE correction (α = 0.05). BPI was implemented using γ = 1 *prior SD*$_θ$. The error bars represent the mean and standard deviation across ten random groups. Horizontal lines indicate the true number of practically significant voxels.

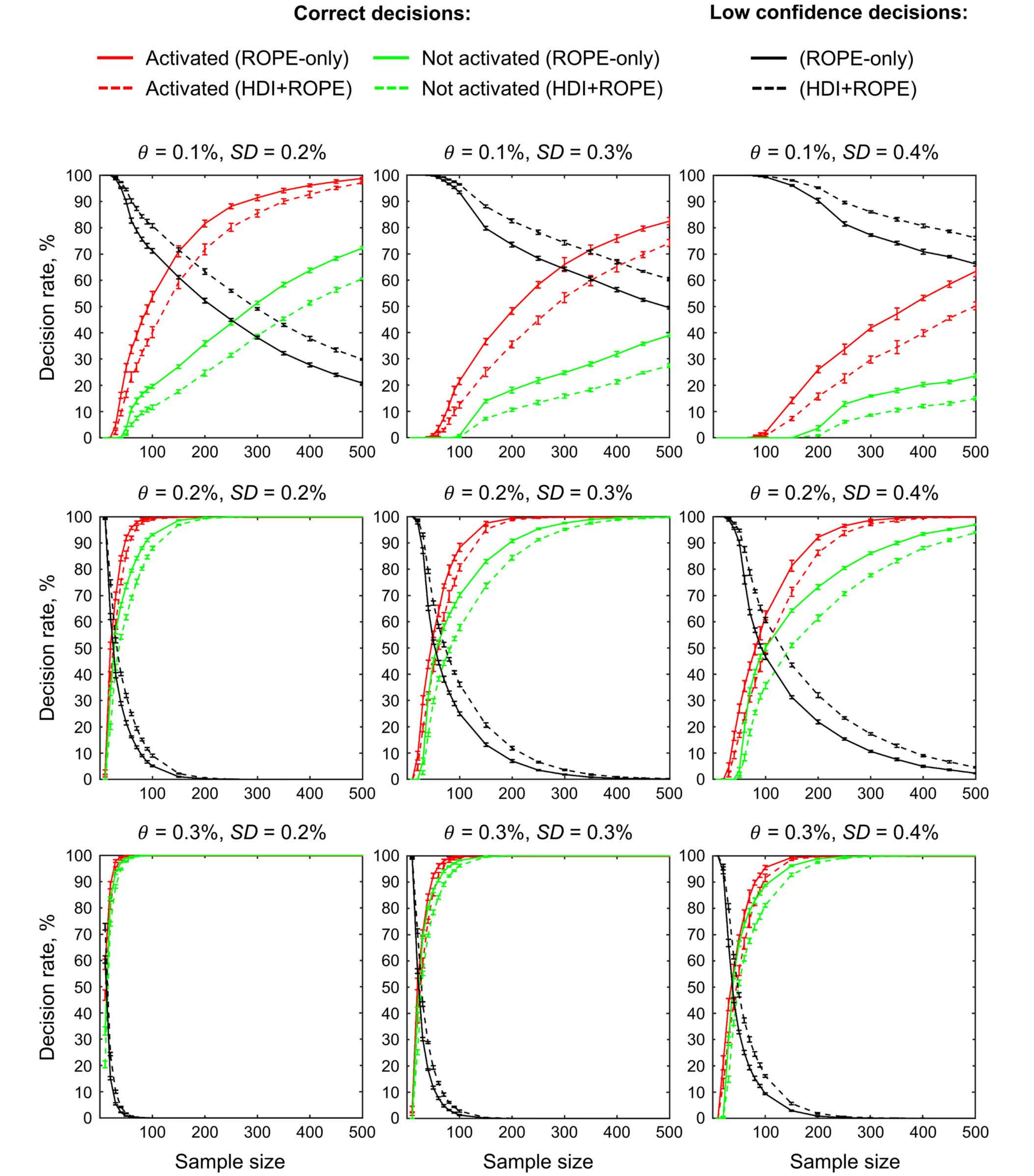

FIGURE 16 | Simulations results for the dependencies of the correct and low confidence decision rates on the sample size. Data were drawn from normal distributions with different mean effect θ and noise *SD*. BPI was implemented using γ = 1 *prior* SD_{θ}. The plots for 'deactivated' voxels closely follow the plots for 'activated' voxels and have therefore been omitted for visualization purposes. The error bars represent the mean and standard deviation across ten random groups.

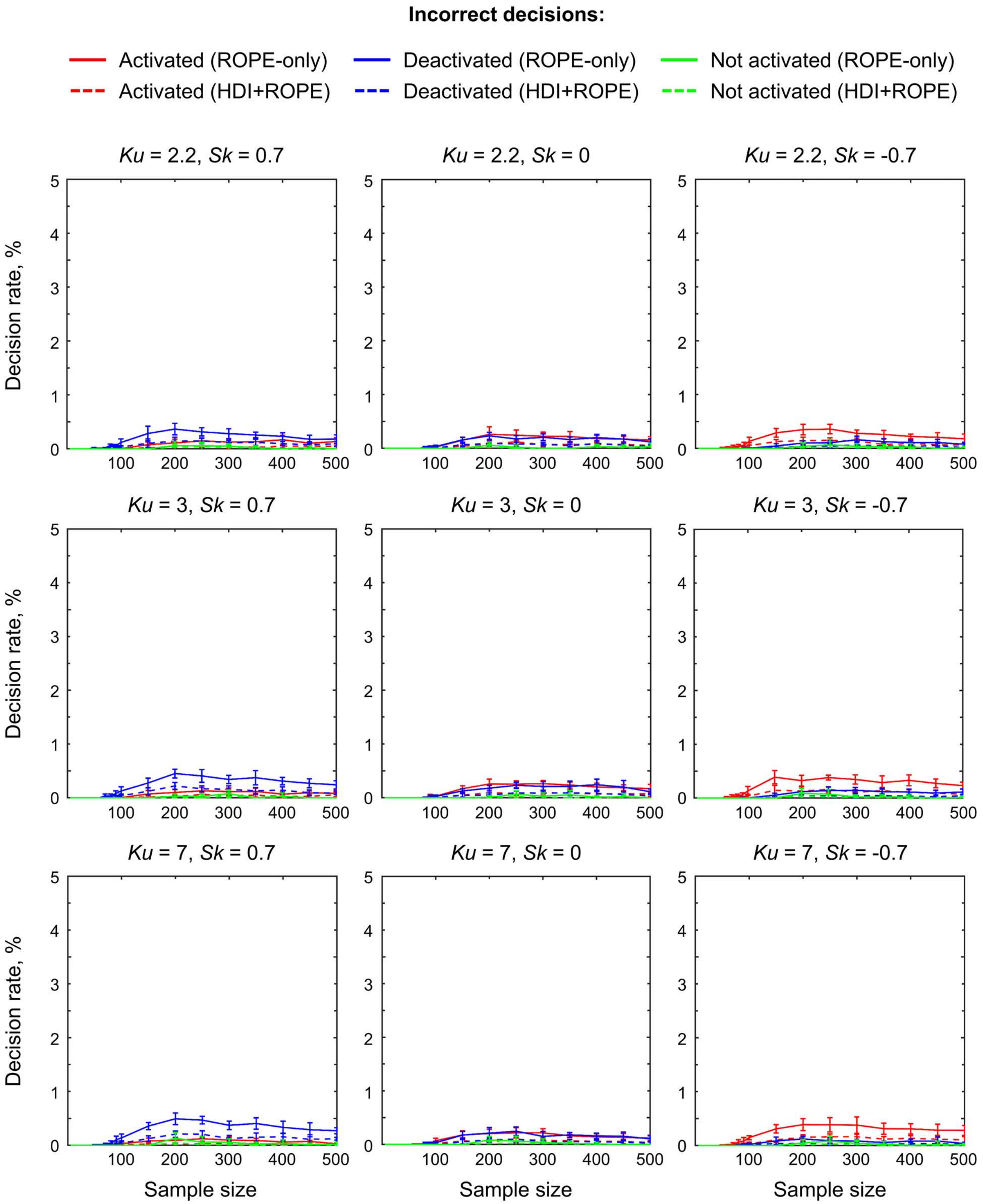

FIGURE 17 | Simulations results for the dependencies of the incorrect decision rate on the sample size. Data were drawn from normal (*Ku* = 3, *Sk* = 0) and non-normal distributions with weak effect and high noise (θ = 0.1%, *SD* = 0.4%). BPI was implemented using γ = 1 *prior* SD_θ. The error bars represent the mean and standard deviation across ten random groups.

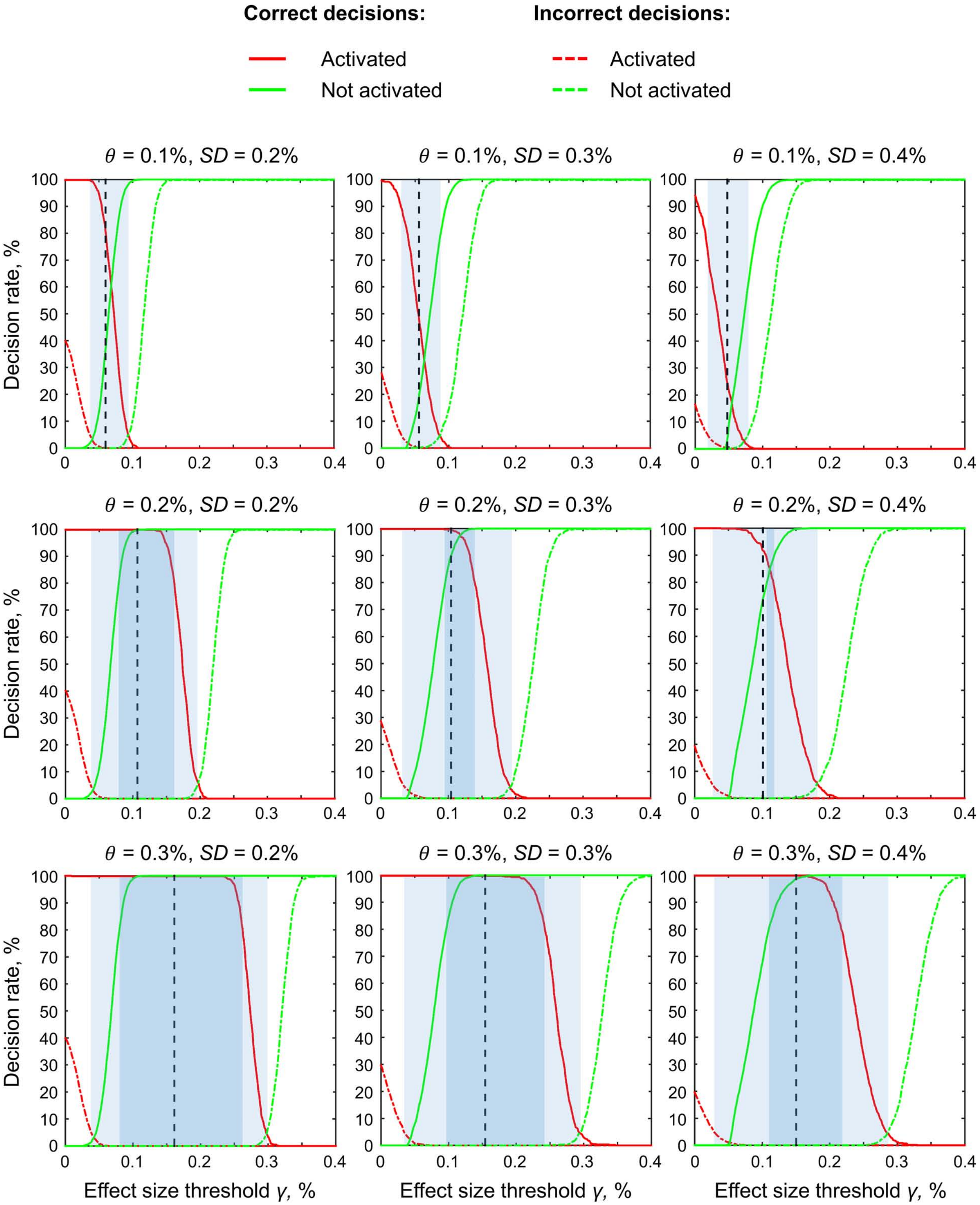

FIGURE 18 | Simulations results for the dependencies of the correct and incorrect decision rates on the ES threshold γ. Data were drawn from normal distributions with different mean effect θ and noise *SD*. Sample size *N* = 200 images, results for one random group. The plots for 'deactivated' voxels closely follow the plots for 'activated' voxels and have therefore been omitted for visualization purposes. Vertical lines indicate the default ES threshold γ = 1 *prior* SD_θ. The light blue areas indicate ES thresholds at which the incorrect decision rates do not exceed 5% for both 'activated' and 'not activated' voxels. The dark blue areas indicate ES thresholds at which the correct decision rates exceed 80% for both 'activated' and 'not activated' voxels.

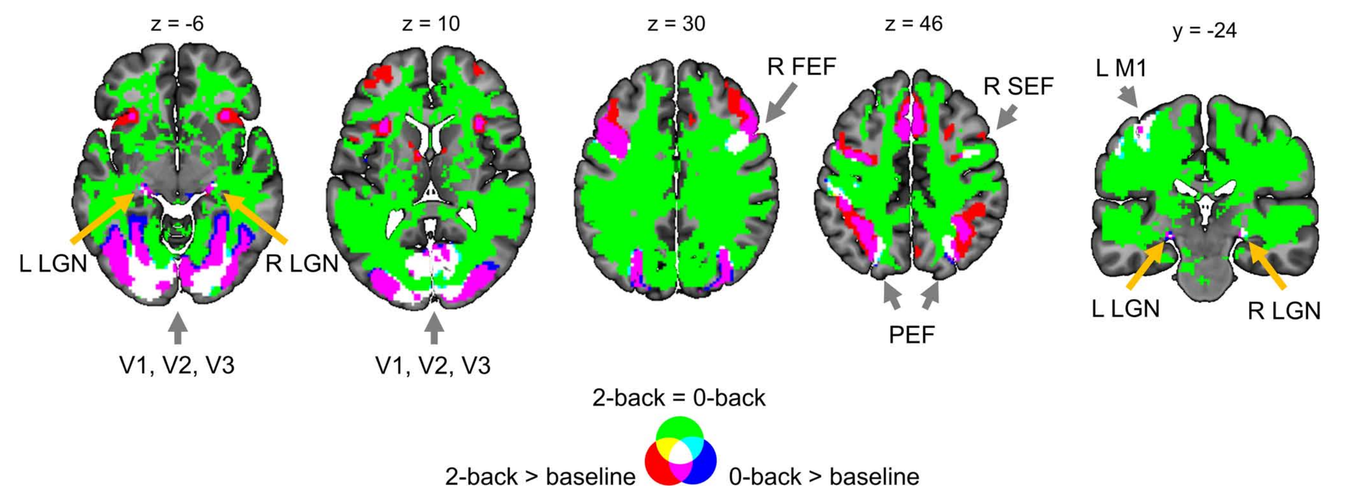

FIGURE 19 | Example of possible application of BPI based on the working memory task. L/R, left/right; V1, V2, V3, primary, secondary, and third visual cortex; FEF, frontal eye field; SEF, superior eye field; PEF, parietal eye field; LGN, lateral geniculate nucleus; M1, primary motor cortex.

reach the plateau and still label the brain area as 'low confidence.' However, even in this case, we can make a definite conclusion: the task design is not sensitive to the effect and should be revised. Empirical Bayes with the 'global shrinkage' prior allows us to monitor the evidence for the alternative or null hypotheses after each participant without special adjustment for multiplicity (Edwards et al., 1963; Berger and Berry, 1988; Wagenmakers, 2007; Rouder, 2014; Kruschke and Liddell, 2017b; Schönbrodt et al., 2017). The optional stopping of the experiment not only allows more freedom in terms of the experimental design, but also saves limited resources and is even more ethically justified in certain cases[6] (Edwards et al., 1963; Wagenmakers, 2007). To strike a balance between analytical flexibility and subjectivity of analysis, one may pre-register hypotheses, models, priors and desired level of evidence to reach without being limited by predefined sample size.

In contrast, frequentist inference depends on the researcher's intention to stop data collection and thus requires a definition of the stopping rule based on *a priori* power analysis. The sequential analysis and optional stopping in frequentist inference inflate the number of false positives and require special multiplicity adjustments. Moreover, even if the *a priori* defined sample size is reached, the researcher can still obtain a non-significant result. In this case, the researcher can follow two controversial paths within the classical NHST framework. Firstly, the sample size could be further increased to force an indecisive result to a decisive conclusion. The problem is that this conclusion would always be against the null hypothesis. Thus, an unbounded increase in the sample size introduces a discrepancy between classical NHST and Bayesian inference, also known as the Jeffreys-Lindley paradox. Secondly, one may argue that high *a priori* power and non-significant results provide evidence for the null hypothesis (see, for example, Cohen, 1990). However, even high *a priori* power and non-significant results do not provide direct evidence for the null hypothesis. In fact, a high-powered non-significant result may arise when the obtained data provide no evidence for the null over the alternative hypothesis, according to Bayesian inference (Dienes and Mclatchie, 2017). This does not mean that power analysis is irrelevant from a Bayesian perspective. Although power analysis is not necessary for Bayesian inference, it can still be used within the Bayesian framework for study planning (Kruschke and Liddell, 2017b). At the same time, power analysis is a critical part of frequentist inference, as it depends on researcher intentions, such as the stopping intention.

The main difficulty with the application of BPI is the need to define the ES threshold. However, the problem of choosing a practically meaningful effect size is not unique to fMRI studies, as it arises in every mature field of science. It should not discourage us from using BPI, as the point-null hypothesis is never true in the soft sciences. From our perspective, there are several ways to address this problem. Firstly, the ES threshold can be chosen based on previously reported effect sizes in studies with a similar design or perform a pilot study to estimate the expected effect size.

Based on the fMRI literature, the largest BOLD responses are evoked by sensory stimulation and vary within 1–5% of the overall mean whole-brain activity. In contrast, BOLD responses induced by cognitive tasks vary within 0.1–0.5% (Friston et al., 2002b; Poldrack et al., 2011; Chen et al., 2017). The results obtained in this study support this notion. Primary sensory effects were >1%, and motor effects were >0.3%. Cognitive effects can be classified into three categories.

(1) 'Strong' effects of 0.2−0.3% (for example, emotion processing in the amygdala, language processing in Broca's area),

[6] This is especially true for PET studies. The BPI method described in this work can also be applied to PET data to reduce the sample size and thus exposure to radioactivity (Svensson et al., 2020).

(2) 'Moderate' effects of 0.1—0.2% (for example, working memory load in DLPFC, social cognition in IPL, response inhibition in IFG/FO),
(3) 'Weak' effects of 0.05–0.1% in contrasts without robust activations (for example, reward processing in the nucleus accumbens, relational processing in DLPFC).

However, choosing the ES threshold based on previous studies can be challenging because fMRI designs become increasingly complex over time, and it can be difficult to find previous experiments reporting unbiased effect size with a similar design. In this case, one can use the ES threshold equal to *one prior SD* of the effect (Friston and Penny, 2003), which can be thought as a neuronal 'background noise level' or a level of activity that is generic to the whole brain (Eickhoff et al., 2008). As empirical and simulation analysis results show, BPI with this ES threshold generally works well for both 'activated/deactivated' and 'not activated' voxel detection. However, it may not be suitable in cases with the weak effects and high noise. In addition, researchers who rely more on the frequentist inference may use the $\gamma(Dice_{max})$ threshold to replicate the results obtained previously with classical NHST and additionally search for 'not activated' and 'low confidence' voxels. Finally, the degree to which the posterior probability is contained within the ROPEs of different widths could be specified or the ROPE maps in which the thresholding sequence is inverted could be calculated. The ROPE maps can be shared in public repositories, such as Neurovault, along with PPMs, and subsequently thresholded by any reasonable ES threshold.

The ability to provide evidence for the null hypothesis may be especially beneficial for clinical neuroimaging. Possible issues that can be resolved using this approach are:

(1) Let the brain activity in certain ROIs due to a neurodegenerative process decrease by more than γ per year on average without any treatment. To prove that a new treatment *effectively protects against neurodegenerative processes*, we can provide evidence that, within 1 year of treatment, brain activity was reduced by less than X%.
(2) Assume that an effective treatment should change the brain activity in certain ROIs by at least X%. Then, we can prove that a new treatment is *practically ineffective* if the activity has changed by less than X%.
(3) Consider two groups of subjects taking a new treatment and a placebo, respectively. Using BPI, we can provide evidence that the result of the new treatment is *does not differ from that of the placebo*.
(4) Consider two groups of subjects taking an old effective treatment and a new treatment. Using BPI, we can provide evidence that the new treatment is *no worse than the old effective treatment*.
(5) Consider a new treatment for a disease that *is not related to brain function*. Using BPI, we can provide evidence that the new treatment *does not have side effects* on brain activity.

CONCLUSION

Herein, a discussion of the use of the Bayesian and frequentist approaches to assess the 'null effects' in fMRI studies was presented. We demonstrated that group-level Bayesian inference may be more intuitive and convenient in practice than frequentist inference. Crucially, Bayesian inference can detect 'activated/deactivated,' 'not activated,' and 'low confidence' voxels using a single decision rule. Moreover, it allows for interim analysis and optional stopping when the obtained sample size is sufficient to make a confident inference. We considered the problem of defining a threshold for the effect size and provided a reference set of typical effect sizes in different fMRI designs. Bayesian inference and assessment of the 'null effects' may be especially beneficial for basic and applied clinical neuroimaging. The developed SPM12-based toolbox with a simple GUI is expected to be useful for the assessment of 'null effects' using BPI.

LIMITATIONS AND FUTURE WORK

Firstly, we did not consider BMI, which is currently mainly used for the analysis of effective connectivity. A promising area of future research would be to compare the advantages of BMI and BPI when analyzing local brain activity. Secondly, the 'global shrinkage' prior must be compared with other possible priors, in particular with priors that take into account the spatial dependency between voxels. Thirdly, we used Bayesian statistics only at the group level. Future studies could consider the advantages of using the Bayesian approach at both the subject and group levels.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Human Connectome Project (https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release) and the UCLA Consortium for Neuropsychiatric Phenomics study (https://openneuro.org/datasets/ds000030/versions/1.0.0). Bayesian parameter inference was performed using the SPM12-based toolbox available at https://github.com/Masharipov/Bayesian_inference.

AUTHOR CONTRIBUTIONS

RM, AK, and MK contributed to conceptualization of the research. MK supervised the project. RM, IK, and YN contributed to statistical analysis and programming. RM and IK performed simulations. MD, DC, and MK acquired funding. All authors contributed to the text of this article and approved the submitted version.

FUNDING

RM, AK, and MK were supported by the Russian Science Foundation Grant #19-18-00454. IK, YN, MD, and DC were supported by the state assignment of the Ministry of Education and Science of Russian Federation (theme number AAAA-A19-119101890066-2). Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Another part of the data were provided by UCLA dataset which was obtained from the OpenfMRI database (its accession number is ds000030) and data collection was funded by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580, RL1MH083268, RL1MH083269, RL1DA024853, RL1MH083270, RL1LM009833, PL1MH083271, and PL1NS062410).

ACKNOWLEDGMENTS

We thank Andrey Ogai for the valuable help with a code for visualization of statistical maps.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.738342/full#supplementary-material

REFERENCES

Acar, F., Seurinck, R., Eickhoff, S. B., and Moerkerke, B. (2018). Assessing robustness against potential publication bias in activation likelihood estimation (ALE) meta-analyses for fMRI. *PLoS One* 13:e0208177. doi: 10.1371/journal.pone.0208177

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., et al. (2018). Quantifying support for the null hypothesis in psychology: an empirical investigation. *Adv. Methods Pract. Psychol. Sci.* 1, 357–366. doi: 10.1177/2515245918773742

Alberton, B. A., Nichols, T. E., Gamba, H. R., and Winkler, A. M. (2020). Multiple testing correction over contrasts for brain imaging. *Neuroimage* 216:116760. doi: 10.1016/j.neuroimage.2020.116760

Altman, D. G., and Bland, J. M. (1995). Statistics notes: absence of evidence is not evidence of absence. *BMJ* 311:485. doi: 10.1136/bmj.311.7003.485

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ* 5:e3544. doi: 10.7717/peerj.3544

Baguley, T. (2009). Standardized or simple effect size: what should be reported? *Br. J. Psychol.* 100, 603–617. doi: 10.1348/000712608x377117

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189. doi: 10.1016/j.neuroimage.2013.05.033

Belia, S., Fidler, F., Williams, J., and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* 10, 389–396. doi: 10.1037/1082-989x.10.4.389

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* 18, 1–32. doi: 10.1214/ss/1056397485

Berger, J. O., and Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *Am. Sci.* 76, 159–165.

Berger, J. O., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence: rejoinder. *J. Am. Stat. Assoc.* 82:135. doi: 10.2307/2289139

Berry, D. (1988). "Multiple comparisons, multiple tests, and data dredging: a Bayesian perspective," in *Bayesian Statistics*, eds J. Bernardo, M. DeGroot, D. Lindley, and A. Smith (Oxford: Oxford University Press), 79–94.

Berry, D. A., and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Inference* 82, 215–227. doi: 10.1016/s0378-3758(99)00044-0

Campbell, H., and Gustafson, P. (2018). Conditional equivalence testing: an alternative remedy for publication bias. *PLoS One* 13:e0195145. doi: 10.1371/journal.pone.0195145

Chen, G., Cox, R. W., Glen, D. R., Rajendra, J. K., Reynolds, R. C., and Taylor, P. A. (2018). A tail of two sides: artificially doubled false positive rates in neuroimaging due to the sidedness choice with t -tests. *Hum. Brain Mapp.* 40, 1037–1043. doi: 10.1002/hbm.24399

Chen, G., Taylor, P. A., and Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *Neuroimage* 147, 952–959. doi: 10.1016/j.neuroimage.2016.09.066

Chen, G., Taylor, P. A., Cox, R. W., and Pessoa, L. (2020). Fighting or embracing multiplicity in neuroimaging? Neighborhood leverage versus global calibration. *Neuroimage* 206:116320. doi: 10.1016/j.neuroimage.2019.116320

Chen, G., Xiao, Y., Taylor, P. A., Rajendra, J. K., Riggins, T., Geng, F., et al. (2019). Handling multiplicity in neuroimaging through Bayesian lenses with multilevel modeling. *Neuroinformatics* 17, 515–545. doi: 10.1007/s12021-018-9409-6

Cohen, J. (1965). "Some statistical issues in psychological research," in *Handbook of Clinical Psychology*, ed. B. B. Wolman (New York, NY: McGraw-Hill), 95–121.

Cohen, J. (1990). Things I have learned (so far). *Am. Psychol.* 45, 1304–1312. doi: 10.1037/0003-066x.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066x.49.12.997

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *Am. Stat.* 20, 18–23. doi: 10.1080/00031305.1966.10479786

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychol. Methods* 2, 161–172. doi: 10.1037/1082-989x.2.2.161

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., et al. (2015). Hidden multiplicity in exploratory multiway ANOVA: prevalence and remedies. *Psychon. Bull. Rev.* 23, 640–647. doi: 10.3758/s13423-015-0913-5

Cremers, H. R., Wager, T. D., and Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLoS One* 12:e0184923. doi: 10.1371/journal.pone.0184923

Cumming, G. (2013). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Dandolo, L. C., and Schwabe, L. (2019). Time-dependent motor memory representations in prefrontal cortex. *Neuroimage* 197, 143–155. doi: 10.1016/j.neuroimage.2019.04.051

David, S. P., Naudet, F., Laude, J., Radua, J., Fusar-Poli, P., Chu, I., et al. (2018). Potential reporting bias in neuroimaging studies of sex differences. *Sci. Rep.* 8:6082. doi: 10.1038/s41598-018-23976-1

de Winter, J. C., and Dodou, D. (2015). A surge ofp-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 3:e733. doi: 10.7717/peerj.733

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Dienes, Z., and Mclatchie, N. (2017). Four reasons to prefer Bayesian analyses over significance testing. *Psychon. Bull. Rev.* 25, 207–218. doi: 10.3758/s13423-017-1266-z

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242. doi: 10.1037/h0044139

Eickhoff, S. B., Grefkes, C., Fink, G. R., and Zilles, K. (2008). Functional lateralization of face, hand, and trunk representation in anatomically defined human somatosensory areas. *Cereb. Cortex* 18, 2820–2830. doi: 10.1093/cercor/bhn039

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Falk, R., and Greenbaum, C. W. (1995). Significance tests die hard. *Theory Psychol.* 5, 75–98. doi: 10.1177/0959354395051004

Feng, C., Forthman, K. L., Kuplicki, R., Yeh, H. W., Stewart, J. L., and Paulus, M. P. (2019). Neighborhood affluence is not associated with positive and negative valence processing in adults with mood and anxiety disorders: a Bayesian inference approach. *Neuroimage Clin.* 22:101738. doi: 10.1016/j.nicl.2019.101738

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., and Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* 20, 1539–1544. doi: 10.1111/j.1523-1739.2006.00525.x

Finch, S., Cumming, G., and Thomason, N. (2001). Colloquium on effect sizes: the roles of editors, textbook authors, and the publication manual. *Educ. Psychol. Meas.* 61, 181–210. doi: 10.1177/0013164401612001

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *Neuroimage* 61, 1300–1310. doi: 10.1016/j.neuroimage.2012.04.018

Friston, K. (2013). Sample size and the fallacies of classical inference. *Neuroimage* 81, 503–504. doi: 10.1016/j.neuroimage.2013.02.057

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: theory. *Neuroimage* 16, 465–483. doi: 10.1006/nimg.2002.1090

Friston, K., Glaser, D., Henson, R., Kiebel, S., Phillips, C., and Ashburner, J. (2002b). Classical and Bayesian inference in neuroimaging: applications. *Neuroimage* 16, 484–512. doi: 10.1006/nimg.2002.1091

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage* 34, 220–234. doi: 10.1016/j.neuroimage.2006.08.035

Friston, K., and Penny, W. (2003). Posterior probability maps and SPMs. *Neuroimage* 19, 1240–1249. doi: 10.1016/s1053-8119(03)00144-7

Friston, K., and Penny, W. (2011). Post hoc Bayesian model selection. *Neuroimage* 56, 2089–2099. doi: 10.1016/j.neuroimage.2011.03.062

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* 5, 189–211. doi: 10.1080/19345747.2011.618213

Genovese, C. R. (2000). A Bayesian time-course model for functional magnetic resonance imaging data: rejoinder. *J. Am. Stat. Assoc.* 95:716. doi: 10.2307/2669451

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15, 870–878. doi: 10.1006/nimg.2001.1037

Gigerenzer, G. (1993). "The superego, the ego, and the id in statistical reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, eds G. Keren and C. Lewis (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 311–339.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127

Gonzalez-Castillo, J., Saad, Z. S., Handwerker, D. A., Inati, S. J., Brenowitz, N., and Bandettini, P. A. (2012). Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci. US.A.* 109, 5487–5492. doi: 10.1073/pnas.1121049109

Goodman, S. (2008). A dirty dozen: twelve P-value misconceptions. *Semin. Hematol.* 45, 135–140. doi: 10.1053/j.seminhematol.2008.04.003

Goodman, S. N. (1993). p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496. doi: 10.1093/oxfordjournals.aje.a116700

Gopalan, R., and Berry, D. A. (1998). Bayesian multiple comparisons using dirichlet process priors. *J. Am. Stat. Assoc.* 93, 1130–1139. doi: 10.1080/01621459.1998.10473774

Gorgolewski, K. J., Durnez, J., and Poldrack, R. A. (2017). Preprocessed consortium for neuropsychiatric phenomics dataset. *F1000Res.* 6:1262. doi: 10.12688/f1000research.11964.2

Greenland, S. (2019). Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *Am. Stat.* 73(Suppl. 1), 106–114. doi: 10.1080/00031305.2018.1529625

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31, 337–350. doi: 10.1007/s10654-016-0149-3

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychol. Bull.* 82, 1–20. doi: 10.1037/h0076157

Gusnard, D. A., and Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nat. Rev. Neurosci.* 2, 685–694. doi: 10.1038/35094500

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *Am. Psychol.* 52, 15–24. doi: 10.1037/0003-066x.52.1.15

Hodges, J. L., and Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. R. Stat. Soc. Ser. B Methodol.* 16, 261–268. doi: 10.1111/j.2517-6161.1954.tb00169.x

Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006). Probability as certainty: dichotomous thinking and the misuse of p values. *Psychon. Bull. Rev.* 13, 1033–1037. doi: 10.3758/bf03213921

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychon. Bull. Rev.* 21, 1157–1164. doi: 10.3758/s13423-013-0572-3

Hubbard, R., and Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am. Stat.* 57, 171–178. doi: 10.1198/0003130031856

Hubbard, R., and Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory Psychol.* 18, 69–88. doi: 10.1177/0959354307086923

Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.* 18, 235–241. doi: 10.1016/j.tics.2014.02.010

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p values? *Am. Stat.* 73(Suppl. 1), 20–25. doi: 10.1080/00031305.2018.1447512

Jeffreys, H. (1939/1948). *Theory of Probability*, 2nd Edn. Oxford: The Clarendon Press.

Jennings, R. G., and Van Horn, J. D. (2012). Publication bias in neuroimaging research: implications for meta-analyses. *Neuroinformatics* 10, 67–80. doi: 10.1007/s12021-011-9125-y

Johansson, T. (2011). Hail the impossible: p-values, evidence, and likelihood. *Scand. J. Psychol.* 52, 113–125. doi: 10.1111/j.1467-9450.2010.00852.x

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 6. New York, NY: John Wiley and Sons, 1–119.

Joyce, K. E., and Hayasaka, S. (2012). Development of PowerMap: a software package for statistical power calculation in neuroimaging studies. *Neuroinformatics* 10, 351–365. doi: 10.1007/s12021-012-9152-3

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educ. Psychol. Meas.* 56, 746–759. doi: 10.1177/0013164496056005002

Knief, U., and Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behav. Res. Methods* 1–15. doi: 10.3758/s13428-021-01587-5

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends Cogn. Sci.* 14, 293–300. doi: 10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 6, 299–312. doi: 10.1177/1745691611406925

Kruschke, J. K., and Liddell, T. M. (2017b). The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* 25, 178–206. doi: 10.3758/s13423-016-1221-4

Kruschke, J. K., and Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychon. Bull. Rev.* 25, 155–177. doi: 10.3758/s13423-017-1272-1

Lakens, D. (2017). Equivalence tests. *Soc. Psychol. Pers. Sci.* 8, 355–362. doi: 10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., and Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *J. Gerontol. Ser. B* 75, 45–57. doi: 10.1093/geronb/gby065

Liao, J. G., Midya, V., and Berg, A. (2019). Connecting Bayes factor and the region of practical equivalence (ROPE) procedure for testing interval null hypothesis. *arXiv* [Preprint] arXiv:1903.03153,

Lindley, D. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, 1st Edn. Cambridge: Cambridge University Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* 44:187. doi: 10.2307/2333251

Lindley, D. V. (1975). The future of statistics: a Bayesian 21st century. *Adv. Appl. Probab.* 7:106. doi: 10.2307/1426315

Lindley, D. V. (1990). The 1988 wald memorial lectures: the present position in Bayesian statistics. *Stat. Sci.* 5, 44–65. doi: 10.1214/ss/1177012253

Magerkurth, J., Mancini, L., Penny, W., Flandin, G., Ashburner, J., Micallef, C., et al. (2015). Objective Bayesian fMRI analysis–a pilot study in different clinical environments. *Front. Neurosci.* 9:168. doi: 10.3389/fnins.2015.00168

Meehl, P. E. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philos. Sci.* 34, 103–115. doi: 10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806–834. doi: 10.1037/0022-006X.46.4.806

Meyners, M. (2012). Equivalence tests – a review. *Food Qual. Prefer.* 26, 231–245. doi: 10.1016/j.foodqual.2012.05.003

Morey, R. D., Hoekstra, R., Rouder, J. N., and Wagenmakers, E. J. (2015). Continued misinterpretation of confidence intervals: response to Miller and Ulrich. *Psychon. Bull. Rev.* 23, 131–140. doi: 10.3758/s13423-015-0955-8

Morey, R. D., and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychol. Methods* 16, 406–419. doi: 10.1037/a0024377

Muller, P., Parmigiani, G., and Rice, K. (2006). "FDR and Bayesian multiple comparisons rules," in *Proceedings of the 8th Valencia International Meeting Bayesian Statistics 8*, eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, et al. (Oxford: Oxford University Press), 366–368.

Mumford, J. A. (2012). A power calculation guide for fMRI studies. *Soc. Cogn. Affect. Neurosci.* 7, 738–742. doi: 10.1093/scan/nss059

Mumford, J. A., and Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* 39, 261–268. doi: 10.1016/j.neuroimage.2007.07.061

Murphy, K. R., and Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model. *J. Appl. Psychol.* 84, 234–248. doi: 10.1037/0021-9010.84.2.234

Murphy, K. R., and Myors, B. (2004). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, 2nd Edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Nichols, T., and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* 12, 419–446. doi: 10.1191/0962280203sm341ra

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage* 62, 811–815. doi: 10.1016/j.neuroimage.2012.04.014

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989x.5.2.241

Penny, W., Flandin, G., and Trujillo-Barreto, N. (2007). Bayesian comparison of spatially regularised general linear models. *Hum. Brain Mapp.* 28, 275–293. doi: 10.1002/hbm.20327

Penny, W., Kiebel, S., and Friston, K. (2003). Variational Bayesian inference for fMRI time series. *Neuroimage* 19, 727–741. doi: 10.1016/s1053-8119(03)00071-5

Penny, W. D., and Ridgway, G. R. (2013). Efficient posterior probability mapping using savage-dickey ratios. *PLoS One* 8:e59655. doi: 10.1371/journal.pone.0059655

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *Neuroimage* 24, 350–362. doi: 10.1016/j.neuroimage.2004.08.034

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* 6:223. doi: 10.3389/fpsyg.2015.00223

Pernet, C. R. (2014). Misconceptions in the use of the general linear model applied to functional MRI: a tutorial for junior neuro-imagers. *Front. Neurosci.* 8:1. doi: 10.3389/fnins.2014.00001

Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., et al. (2016). A phenome-wide examination of neural and cognitive function. *Sci. Data* 3:160110. doi: 10.1038/sdata.2016.110

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., et al. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18, 115–126. doi: 10.1038/nrn.2016.167

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge: Cambridge University Press.

Poline, J. B., and Brett, M. (2012). The general linear model and fMRI: does love last forever? *Neuroimage* 62, 871–880. doi: 10.1016/j.neuroimage.2012.01.133

Pollard, P., and Richardson, J. T. (1987). On the probability of making type I errors. *Psychol. Bull.* 102, 159–163. doi: 10.1037/0033-2909.102.1.159

Raichle, M. E., and Gusnard, D. A. (2002). Appraising the brain's energy budget. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10237–10239. doi: 10.1073/pnas.172399499

Reimold, M., Slifstein, M., Heinz, A., Mueller-Schauenburg, W., and Bares, R. (2005). Effect of spatial smoothing on t-Maps: arguments for going back from t-Maps to masked contrast images. *J. Cereb. Blood Flow Metab.* 26, 751–759. doi: 10.1038/sj.jcbfm.9600231

Rogers, J. L., Howard, K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychol. Bull.* 113, 553–565. doi: 10.1037/0033-2909.113.3.553

Rosa, M., Friston, K., and Penny, W. (2012). Post-hoc selection of dynamic causal models. *J. Neurosci. Methods* 208, 66–78. doi: 10.1016/j.jneumeth.2012.04.013

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Rouder, J. N. (2014). Optional stopping: no problem for Bayesians. *Psychon. Bull. Rev.* 21, 301–308. doi: 10.3758/s13423-014-0595-4

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

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *Am. Stat.* 40:313. doi: 10.2307/2684616

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: CRC Press.

Samartsidis, P., Montagna, S., Laird, A. R., Fox, P. T., Johnson, T. D., and Nichols, T. E. (2020). Estimating the prevalence of missing experiments in a neuroimaging meta-analysis. *Res. Synth. Methods* 11, 866–883. doi: 10.1002/jrsm.1448

Schatz, P., Jay, K., McComb, J., and McLaughlin, J. (2005). Misuse of statistical tests in publications. *Arch. Clin. Neuropsychol.* 20, 1053–1059. doi: 10.1016/j.acn.2005.06.006

Schneider, J. W. (2014). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102, 411–432. doi: 10.1007/s11192-014-1251-5

Schneider, J. W. (2018). NHST is still logically flawed. *Scientometrics* 115, 627–635. doi: 10.1007/s11192-018-2655-4

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychol. Methods* 22, 322–339. doi: 10.1037/met0000061

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* 15, 657–680. doi: 10.1007/bf01068419

Schwartzman, A., Dougherty, R., Lee, J., Ghahremani, D., and Taylor, J. (2009). Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage* 44, 71–82. doi: 10.1016/j.neuroimage.2008.04.182

Serlin, R. C., and Lapsley, D. K. (1985). Rationality in psychological research: the good-enough principle. *Am. Psychol.* 40, 73–83. doi: 10.1037/0003-066x.40.1.73

Serlin, R. C., and Lapsley, D. K. (1993). "Rational appraisal of psychological research and the good-enough principle," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, eds G. Keren and C. Lewis (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 199–228.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611. doi: 10.1093/biomet/52.3-4.591

Sjölander, A., and Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *Eur. J. Epidemiol.* 34, 809–821. doi: 10.1007/s10654-019-00517-2

Smith, S. M., and Nichols, T. E. (2018). Statistical challenges in "big data" human neuroimaging. *Neuron* 97, 263–268. doi: 10.1016/j.neuron.2017.12.018

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31, 2013–2035. doi: 10.1214/aos/1074290335

Streiner, D. L. (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. *Am. J. Clin. Nutr.* 102, 721–728. doi: 10.3945/ajcn.115.113548

Svensson, J., Schain, M., Knudsen, G. M., Ogden, T., and Plavén-Sigray, P. (2020). Early stopping in clinical PET studies: how to reduce expense and exposure. *MedRxiv* [Preprint] doi: 10.1101/2020.09.13.20192856

Szucs, D., and Ioannidis, J. P. (2020). Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* 221:117164. doi: 10.1016/j.neuroimage.2020.117164

Szucs, D., and Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Front. Hum. Neurosci.* 11:390. doi: 10.3389/fnhum.2017.00390

Turkheimer, F. E., Aston, J. A. D., and Cunningham, V. J. (2004). On the logic of hypothesis testing in functional imaging. *Eur. J. Nuclear Med. Mol. Imaging* 31, 725–732. doi: 10.1007/s00259-003-1387-7

UIudag, K., Müller-Bierl, B., and Ugurbil, K. (2009). An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *Neuroimage* 47:S56. doi: 10.1016/s1053-8119(09)70204-6

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/bf03194105

Wagenmakers, E. J., Lee, M., Lodewyckx, T., and Iverson, G. J. (2008). "Bayesian versus Frequentist inference," in *Bayesian Evaluation of Informative Hypotheses. Statistics for Social and Behavioral Sciences*, eds H. Hoijtink, I. Klugkist, and P. A. Boelen (New York, NY: Springer), 181–207.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cogn. Psychol.* 60, 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). "The need for Bayesian Hypothesis testing in psychological science," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld and I. D. Waldman (Hoboken, NJ: Wiley Blackwell), 123–138.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *Am. Stat.* 70, 129–133. doi: 10.1080/00031305.2016.1154108

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd Edn. Milton Park: Taylor & Francis.

Westfall, P., Johnson, W. O., and Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 419–427. doi: 10.1093/biomet/84.2.419

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *J. Pharm. Sci.* 61, 1340–1341. doi: 10.1002/jps.2600610845

Woo, C. W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage* 91, 412–419. doi: 10.1016/j.neuroimage.2013.12.058

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage* 21, 1732–1747. doi: 10.1016/j.neuroimage.2003.12.023

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage* 45, S173–S186. doi: 10.1016/j.neuroimage.2008.10.055

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670. doi: 10.1038/nmeth.1635

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.