Bayesian Meta-analysis of fMRI Image Data

Hyemin Han

University of Alabama

Joonsuk Park

The Ohio State University

Author Note

Correspondence concerning this manuscript should be addressed to Hyemin Han,
Educational Psychology Program, University of Alabama, Box 870231, Tuscaloosa, AL
35487-0231, USA. Tel: 1-205-348-0746. Fax: 1-205-348-2161.

Abstract

We composed an R-based script for Image-based Bayesian random-effect meta-analysis of previous fMRI studies. It meta-analyzes second-level test results of the studies and calculates Bayes Factors indicating whether the effect in each voxel is significantly different from zero. We compared results from Bayesian and classical meta-analyses by examining the overlap between the result from each method and that created by NeuroSynth as the target. As an example, we analyzed previous fMRI studies focusing on working memory extracted from NeuroSynth. The result from our Bayesian method showed a greater overlap than the classical method. In addition, Bayes Factors proved a better way to examine whether the evidence supported hypotheses than p-values. Given these, Bayesian meta-analysis provides neuroscientists with a better meta-analysis method for fMRI studies given the improved overlap with the NeuroSynth result and the practical and epistemological value of Bayes Factors that can directly test presence of an effect.

Keywords: fMRI, Meta-analysis, p-value, Bayesian inference, Bayes Factors, Bayesian random-effect meta-analysis

Bayesian Meta-analysis of fMRI Image Data

Introduction

Meta-analysis has been regarded as a feasible way to address limitations in functional neuroimaging studies, particularly those associated with weak statistical power, which has been a source of serious concerns in cognitive neuroscience (Lieberman & Cunningham, 2009). Neuroscientists, psychologists, and statisticians have collaborated to develop statistical methods for meta-analysis of fMRI studies. Many of them, such as the ALE (Activation Likelihood Estimation), use coordinate-based meta-analysis (CBMA) that estimates the locations and sizes of activation clusters from coordinates of activation foci reported in published articles (Salimi-Khorshidi, Smith, Keltner, Wager, & Nichols, 2009). Although CBMA can be conveniently performed by end-users because it only requires coordinates information reported in articles, it comes with several notable limitations (Salimi-Khorshidi et al., 2009). Particularly, because CBMA estimates activation clusters from coordinates reported in published articles, instead of real image data, loss of important information, such as the actual formation and shape of activation clusters, could be a significant issue. In addition, CBMA is known to be sensitive to reported coordinates information, choice of thresholding scheme, and the type of reporting method employed in each individual study, which limits its reliability and validity.

Image-based meta-analysis (IBMA) has also been recommended by neuroimaging researchers (Salimi-Khorshidi et al., 2009). IBMA is similar to higher-level analysis of statistics image files produced by fMRI studies; it meta-analyzes the reported statistics value in each voxel in previous studies, instead of finding estimated activation clusters as CBMA does. In fact, a previous study reported that IBMA showed a better performance compared to CBMA (Salimi-Khorshidi et al., 2009). Although the availability of shared fMRI images has been an issue for IMBA, recent efforts to establish platforms for image and data sharing, such as Open Science Framework (https://osf.io), NeuroVault (https://neurovault.org), OpenfMRI (https://openfmri.org), may improve feasibility of

IMBA among researchers.

Although the aforementioned meta-analytic approaches have been widely utilized in neuroscience, some statistical issues related to interpreting results should be considered. To date, the majority of meta-analysis methods, including both CBMA and IBMA, have relied on classical or frequentist inference procedures, which focus on thresholding based on p-values, while interpreting results. Unfortunately, over-reliance on such classical procedures would lead us to serious statistical issues (Wagenmakers, 2007). First, p-values themselves do not inform us anything about whether the data supports our hypothesis (e.g., the presence of significant effect in a certain voxel in a certain task condition). In fact, 'a p-value, the probability that one will observe values of the test statistic that are as extreme or more extreme than is actually observed does not directly quantify the likelihood that H_0 or H_1 is correct (p. 3) (Han & Park, 2018).' Second, because statistical testing in neuroimaging studies tends to involve comparisons in more than several hundred thousand voxels, inflated false-positive rates can also be a serious concern (Lieberman & Cunningham, 2009). In fact, previous research has indicated that careless utilization of widely-used analytic tools may cause false positive rates to be higher than the nominal p <.05 (Eklund, Nichols, & Knutsson, 2016), resulting in erroneous reports with Type I errors.

Here we propose that Bayesian hypothesis testing can address the aforementioned issues (Han & Park, 2018). Before discussing the associated benefits, a brief introduction about the 'Bayesian' way of testing hypotheses is in order. Instead of computing a p-value, Bayesians compute what is called the 'Bayes Factor (BF)' (Kass & Raftery, 1995). A BF is defined as the ratio of two marginal probabilities, say $P(D|H_0)/P(D|H_1)$, which is the ratio of probabilities of observing the data at hand under the null and the alternative. Based on whether the BF exceeds a pre-specified threshold, say 3, one accepts the null or the alternative (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

Hypothesis testing via BFs has several advantages when compared to that using

p-values; here we discuss three of them which we deem are most relevant. First, BFs are epistemologically preferable to *p*-values. Their interpretation is more straightforward than that of *p*-values because BFs deal with the probability of observing the dataset at hand, not that of observing a dataset *as extreme or more extreme* than the observed one. This feature of *p*-values has consistently led researchers to mistakenly believe that they were the probability that the null hypothesis is true (Gigerenzer, 2004). Bayesian hypothesis testing directly deals with such a type of probability.

Second, using a BF allows one to actually *accept* the null hypothesis. This contrasts the Bayesian approach to the classical one in a stark way. A strict reading of the classical statistical principles tells us that one cannot literally *accept* the null; one can only 'fail to reject' the null. Although it could be argued that this assertion is false because the decision-theoretic approach, or Neyman-Pearson school of frequentism, allows one to accept the null, it requires one to specify the utility, or loss, function to be incorporated in the decision process, which might be hard to do due to the uncertainty of relative costs in practice. Also, because the decision-theoretic approach is, lamentably, usually intermingled with the Fisherian approach that does not allow one to accept the null, and it is rarely the case that the pure decision theoretic approach is used in practice, we do not consider such a possibility seriously. Despite the aforementioned limitations of classical inference, it is very common for researchers to conclude that there is 'no difference' between experimental groups based on a $p > .05$ (Gigerenzer, 2004). This cannot be justified in the classical framework and is a questionable practice. When it comes to neuroscience, this means that one cannot conclude that a voxel's activation is strictly equal to zero when using *p*-values. In contrast, because one directly compares the probabilities of the data under the null and the alternative, one can 'accept' the null when the BF, $P(D|H_0)/P(D|H_1)$, is sufficiently large. This is an attractive feature of BFs when the null hypothesis is theoretically meaningful, as in the case of fMRI studies where the null says that a voxel is not activated. By using BFs, one can legitimately accept the null, which cannot be done in the frequentist

framework.

Third, Bayesian inference has the potential to addresses concerns about multiple comparison and inflated false positive rates. In our approach, specifically, this is achieved by setting an appropriate threshold for each Bayes factor to be considered as sufficient to accept the alternative hypothesis. We intentionally set the threshold high to address the problem of inflated Type I error rate, which is similar in spirit to the case of Bonferroni correction. This is, in essence, the same approach advocated by traditional Bayesians (Scott & Berger, 2006; Westfall, Johnson, & Utts, 1997). They have proposed to calibrate the prior probability of the alternative hypothesis for each test such that the overall probability that the null is true in all the tests is controlled at some level, specified by the analyst. That is, they adjust the priors to counter the multiplicity problem. However, adjusting the prior is equivalent to adjusting the BF because the posterior odds, which is ultimately used for deciding whether or not to reject the null, is the product of the prior odds and the BF; one can adjust whichever to have the same effect on the resulting posterior odds. Furthermore, instead of adjusting the BF for a fixed threshold, we adjust the threshold for the same BF; contrary to the case of priors, which can be adjusted at the researchers' will, the BF cannot be arbitrarily changed. By changing the threshold for rejecting the null, however, we have the same effect on the result of testing.

Our proposed strategy has been shown to work successfully in practice. For instance, a previous study that compared results from Bayesian versus frequentist second-level fMRI analyses using SPM 12 reported that the former with BF thresholds based on statistical guidelines exhibited better sensitivity and selectivity than the latter that had used frequentist multiple comparison methods, such as voxelwise and clusterwise familywise error correction (Han & Park, 2018). This example shows that Bayesian inference can remedy the problems of inflated false positive rates and multiple comparison in the field of neuroscience.

A possible objection to the use of BF as an alternative to *p*-values is that it still

requires one to set an arbitrary threshold, which seems to be a bad idea. An answer to this criticism is that BF are still preferable to *p*-values in terms of interpretability, as pointed out earlier. At least, by replacing *p*-values with BF, one can quantify the relative evidence in favor of the null and the alternative. In addition, it has been reported that BF are less in favor of the alternative, thereby generating less type I errors (Wetzels et al., 2011), which is also reflected in the better sensitivity and selectivity (Han & Park, 2018).

In addition, it could be argued that posterior probabilities should be used instead of BFs because posteriors are the key quantities in Bayesian inferences. Although we basically agree with the general principle, we recommend using BFs for the following reasons. First, posterior probabilites about the model parameters do not allow one to directly compare the null and the alternative and to quantify the relative evidence provided by the data. Second, BFs do not depend on the model prior that is often difficult to specify. Relatedly, BFs are directly linked to posterior probabilities in case of uniform priors. The use of an equal prior as a reflection of the attitude towards scientific objectivity has been championed (Berger et al., 2006; Rouder et al., 2009). In that case, a Bayes factor is equal to the ratio of marginal posterior probabilities of the hypotheses, or the posterior odds, due to the following equality:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \times \frac{P(H_0)}{P(H_1)}$$

where D denotes the data. Thus, computing a BF is equivalent to obtaining the posterior odds.

In fact, there have been previous studies that applied Bayesian meta-analysis of fMRI data (Kang, Johnson, Nichols, & Wager, 2011; Samartsidis et al., 2018; Yue, Lindquist, & Loh, 2012). They might be able to address the aforementioned issues associated with classical inference; moreover, given that the aforementioned previous studies performed Bayesian meta-analysis, so they can be benefited by the aforementioned statistical advantages of Bayesian inference. However, because they performed CBMA, not IBMA, so they could also suffer the limitations of CBMA. Thus, it would be necessary to explore a

way to implement Bayesian statistics-applied IBMA to address the possible limitations of meta-analyses that have been conducted in the previous studies.

In the present study, we wrote an R script to perform image-based Bayesian meta-analysis of fMRI studies. In addition, as a preliminary exploration, we evaluated its performance by examining the outcome from our Bayesian meta-analysis and that from the large-scale meta-analysis of web-based database, NeuroSynth (http://www.neurosynth.org). We also compared the performance of Bayesian meta-analysis with that of classical meta-analysis. In this processes, we meta-analyzed fMRI statistics image files shared by a web-based image repository. In particular, we focused on the topic of 'working memory' because of the availability of shared image files.

Methods

Materials

We collected images containing results from group-level analysis in previous fMRI studies via NeuroVault. We selected this image repository because NeuroVault aims to collect statistics images resultant from fMRI analysis instead of raw images. We used a keyword, 'working memory,' to search for image collections including analysis results focusing on working memory. As a result, twenty collections were found (as of May 23, 2018). Among them, we downloaded six image files in NIfTI format from five collections. All of these images were demonstrating results from the comparisons between working memory versus non-working memory task condisions. Our image selection process and rationale are presented in Figure S1, and the list of meta-analyzed images is reported in Table S1.

All downloaded images were transformed into MNI152 space (91x109x91). Because all image files reported either T- or Z-statistic, we converted all values into standardized effect sizes in the format of Fisher's z with reported sample sizes. We calculated an effect size indicator, Pearson's r from a given T- or Z-statistic value. Once T- or Z-statistic was

converted into Pearson's r, r was standardized by using a formula as follows:

$$z := \frac{1}{2} ln(\frac{1+r}{1-r}) = arctanh(r)$$

Then, the standard error (se) of each study was calculated as follows:

$$se = \frac{1}{\sqrt{N-3}}$$

This se value enables us to include an important study characteristic, the sample size of a specific study, in the meta-analysis model. If the sample size of a specific study is small, the 95% confidence interval of the effect of this study becomes wide. On the other hand, the 95% confidence interval of the effect of a large study becomes narrow in a meta-analysis model. As a result, in the meta-analysis model, the larger study yields an estimate with the higher precision, while the smaller study yields an estimate with the lower precision. The precision of each individual study determines its impact on the final meta-analysis result (Borenstein, Hedges, Higgins, & Rothstein, 2010).

We used processed image files containing the Fisher's z value of each voxel for further Bayesian and classical meta-analysis.

fMRI meta-analysis

Bayesian meta-analysis. Once all six images were converted into Fisher's z images, they were meta-analyzed with our customized R scripts (related source codes and data files are shared via GitHub, https://github.com/xxelloss/BFMeta). Readers may analyze their own dataset with a script shared in the repository. If readers intend to run the script, there are a couple of steps that they have to complete before running the script. First, NIfTI images containing z or t-statistics produced by each individual study should be prepared; the coordinates system used in the NIfTI files should be MNI coordinates (91 x 109 x 91), or the sections in the script dealing with the coordinates system should be modified. Second, the readers should create (or modify) list.csv file that contains the filename, sample size, and type of statistics of each study.

For Bayesian meta-analysis, we used R package named metaBMA (Gronau et al., 2017; Heck, Gronau, & Wagenmakers, 2017). A vector containing converted effect size values in previous studies in a specific voxel was created. We entered this vector and a list of the standard error of each study to a meta-analysis model. In this process, we used meta_random function that implemented random-effect meta-analysis to address potential between-study heterogeneity (Gronau et al., 2017). Because Bayesian inference requires prior distributions as inputs, we used these priors for main effect, r, and standard deviation of study effects, τ, that were suggested by Gronau et al.(Gronau et al., 2017):

$$r \sim Cauchy(\frac{1}{\sqrt{2}})$$

$$\tau \sim \beta(1, 2)$$

Then, the prior probability distribution of effect was updated with data through iterations. In these processes, the Markov chain Monte Carlo (MCMC) method was employed. After 4,000 burnin iterations, 10,000 samples were used for iterative processes in the MCMC simulation. Once the iterative processes were completed, three output values, a BF, and estimated mean and median effect size, for each voxel were calculated. We created three NIfTI image files containing the aforementioned three values. Mean and median effect size values were converted into Pearson's r for interpretation as follows:

$$r = \frac{exp(2z) - 1}{exp(2z) + 1} = tanh(z)$$

For further evaluations, we thresholded resultant images based on BF. Following Kass and Raftery's guidelines (Kass & Raftery, 1995), we used 2logBF values. Three thresholded mask images were created using three different BF thresholds, 2logBF = 2, 2logBF = 6, and 2logBF = 10, represent the presence of positive, strong, and very strong evidence supporting effects, respectively.

Because our sample size ($n = 6$) was small, we examine the robustness of the Cauchy prior. This robustness check was performed to test whether the result was enormously

sensitive to the change in the prior distribution or the reported result was merely resulted from peculiarity. Following the prior robustness check method implemented in JASP (Quintana & Williams, 2018), we plotted the resultant BF value in a specific voxel while changing the parameter of the Cauchy distribution from .01 to 1.5. In addition, to examine whether the type of the prior distribution significantly influenced the posterior distribution, we performed the same robustness check with a normal distribution. We also plotted the calculated BF values while changing the standard deviation parameter of the normal distribution from .01 to 5.0 (Mean = .00).

Classical meta-analysis.   To compare performance of Bayesian and classical inference in meta-analysis, we performed classical meta-analysis with the same images with our customized R scripts (source codes are available via our GitHub repository). The same effect size images and standard error values used in Bayesian meta-analysis were reused. All effect size values were transformed into Fisher's z. Instead of meta_random function in metaBMA package, we performed classical random-effect meta-analysis with rma function implemented in metafor package (Viechtbauer, 2010). In this process the Maximum Likelihood approach was employed for estimation. As outputs, images containing the z-score and estimated effect size in Pearson's r of each voxel were produced.

For evaluations, we thresholded z-score images based on p-values. We created two thresholded mask images using different voxelwise thresholds. Following Salimi-Khorshidi et al. (2009), we decided to use $p < .05$ after correcting for the false positive discovery rate (FDR) as a primary threshold. Given that we are mainly interested in the way how end users perform meta-analysis, the aforementioned de facto primary threshold for classical inference in IBMA was employed in the present study. For exploratory purposes, we applied more stringent thresholds, $p < .01$, $p < .005$, and $p < .001$ after a FDR correction. The FDR correction was performed with xjview (Cui, Li, & Song, 2015).

Performance evaluation

We assessed the performance of each meta-analysis method by utilizing the result from large-scale synthesis as a standard. We used a result from NeuroSynth that meta-analyzes 3,489 published neuroimaging studies (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). We particularly focused on results from reverse inference because reverse inference identifies 'the probability that a term (task) would occur in an article given the presence of activation in a particular brain region (p. 667 (Yarkoni et al., 2011)),' P (term|activation). On the other hand, the traditional inference method, forward inference quantifies 'the probability that there would be activation in specific brain regions given the presence of a particular term (task), P (activation|term) (p. 667 (Yarkoni et al., 2011)).' Because we were mainly interested in the specific neural activity pattern of the task condition of our interest, working memory, we decided to use reverse inference as the primary source of evaluation. However, for exploratory purposes, we also compared results from meta-analyses and the result from NeuroSynth forward inference as well.

We extracted a NIfTI image containing the result of reverse inference by using a term, 'working memory,' in NeuroSynth webpage (http://neurosynth.org/analyses/terms/working%20memory/). The automatically created synthesis image was downloaded for further evaluations. Then, we calculated an overlap index that indicates whether results from different meta-analysis methods well overlapped with the synthesis image by using a formula as follows:

$$I_{ovl} = \frac{2\frac{V_{ovl}V_{ovl}}{V_{meta}V_{synth}}}{\frac{V_{ovl}}{V_{meta}} + \frac{V_{ovl}}{V_{synth}}}$$

where V_{ovl} is the number of voxels showing significant activity in both images, and V_{meta} and V_{synth} are the number of voxels showing significant activity in the result of our meta-analysis and that from NeuroSynth, respectively (Han & Glenn, 2018). This overlap index was employed in the present study because it does take into account both false positive voxels, voxels reported to be active in the meta-analysis result image but are not

in fact active in the standard image, and false negative voxels, voxels reported to be inactive in the meta-analysis image but are in fact active in the standard image. In our study, we assumed that the higher overlap index indicates the higher performance of a meta-analysis method.

In addition to the overlap index, we also used the ratio of the number of voxels showing results inconsistent with the NeuroSynth standard to that of voxels showing results consistent with the NeuroSynth standard. This ratio can be calculated as follows:

$$R_{errortocorrect} = \frac{V_{falsealarm} + V_{miss}}{V_{ovl}}$$

$V_{false\ alarm}$ is the number of voxels that were found to be significantly activated from our meta-analysis, but were not actually significantly activated in a NeuroSynth image. On the other hand, V_{miss} represents the number of voxels that showed significant activity in our meta-analysis, but were not in fact significantly active in a NeuroSynth image. We use this ratio to examine to which extent a specific meta-analysis method produced results inconsistent with the NeuroSynth result. We assumed that the higher $R_{error\ to\ correct}$ indicates the worse performance of a specific meta-analysis method.

In addition to the reverse and forward inference NeuroSynth results, we used one more meta-analysis image from a previous study as a standard for comparisons (Rottschy et al., 2012). The previous study meta-analyzed 189 fMRI experiments focusing on working memory tasks with ALE, a CBMA algorithm, to examine the common neural correlates of working memory processing. We downloaded the result of the meta-analysis from NeuroVault (https://neurovault.org/images/12186/) in the format of NIfTI. The previous study used a cluster-level corrected threshold of $p < .05$ with a cluster-forming threshold at voxel-level $p < .001$.

Performance evaluation with simulation data

We conducted performance evaluation with an additional dataset that contains both original images ($n = 6$) and simulation images to increase the overall sample size; in

addition, we also intended to examine whether Bayesian analysis is more robust when image files that contain a noise are analyzed. To produce additional image files, we added the random noise to each original image file. We added the random noise ($SD = 5\%$ of the mean signal strength in each original image) while creating new image files Han and Park (2018). As a result, for this additional performance evaluation, we used a total of twelve image files, six original and six newly created image files from simulation that contain the random noise.

These twelve files were meta-analyzed with Bayesian and classical IBMA methods by using our customized R scripts as explained in prior sections. The same meta-analysis procedures, and BF (2logBF = 2, 6, and 10) and p-value thresholds ($p < .05, .01, .005,$ and $.001$ after FDR corrected) were applied. For performance evaluation, we calculated I_{Ovl} and $R_{error\ to\ correct}$ in all cases.

Results

Results from Bayesian meta-analysis are presented in Figures 1a-c. First, as shown in Figure 1a, applying the more stringent 2logBF threshold resulted in the less survived voxels after thresholding. Voxels demonstrating significant activation in NeuroSynth are also presented in the same figure for reference. Mean and median effect size maps are presented in Figures 1b and c, respectively. Figure 1d demonstrates voxels survived p-value thresholding from classical meta-analysis. Similar to Bayesian meta-analysis, applying the application of the more stringent threshold (.01 vs. .05) resulted in the less survived voxels.

We evaluated the performance of each meta-analysis method by comparing meta-analysis outcomes and results from NeuroSynth reverse inference, NeuroSynth forward inference, and previous CBMA ALE meta-analysis. NeuroSynth maps created by using a keyword, 'working memory,' at NeuroSynth were compared with thresholded images from our meta-analyses. In addition, the same comparison was performed with the previous CBMA ALE meta-analysis result. We quantified the degree of overlap in terms of

an overlap index, and results are reported in Figure 2a. In addition, $R_{error\ to\ correct}$ values are presented in Figure 2b. As shown in the figures, Bayesian meta-analysis in general showed greater overlap indices and lower $R_{error\ to\ correct}$ values compared with classical meta-analysis. The overlap index was highest when $2logBF = 6$ that indicates the presence of strong evidence was used for thresholding. One exception was the comparison with NeuroSynth forward inference. Once the more stringent thresholds (e.g., $p < .01$, .005, or .001) were applied, classical inference reported slightly better outcomes compared with Bayesian inference even when $2logBF = 6$ was used for the Bayesian inference threshold.

The prior robustness check was performed with a voxel value at (0, -6, 0). We selected this specific voxel because it was closest voxel to the center, (0, 0, 0), with the calculated BF greater than $2logBF = 6$ threshold. The original calculated $2logBF$ value at this voxel was 7.06. Figure 3 shows results from the prior robustness checks. When we changed the parameter of the Cauchy distribution prior, in all cases, the calculated BF was greater than $2logBF = 2$ threshold representing the presence of positive evidence supporting a significant effect; in the most cases, the BF exceeded $2logBF = 6$ threshold suggesting the presence of strong evidence (see Figure 3a). Similar results were reported when we tested the normal distribution prior. Except for extreme cases, when the distribution was extremely sharp ($SD < .10$), the calculated BF was greater than either $2logBF = 2$ or $2logBF = 6$ threshold (see Figure 3b).

Finally, results from the performance analysis of the additional dataset are presented in Figures S1 and S2. Similar to the results from meta-analyses of six images, Bayesian IBMA demonstrated better performance compared with classical IBMA when twelve images were meta-analyzed. Interestingly, when $2logBF = 6$ or $2logBF = 10$ threshold was applied, Bayesian IBMA outperformed classical IBMA regardless of which p-value threshold was applied and which meta-analysis image was used as a standard for comparison. Even in the case of the comparison with the forward inference image, Bayesian IBMA reported better performance indicators compared with classical IBMA

when 2logBF = 6 or 2logBF = 10 BF threshold was utilized.

Discussion

We developed an R program performing image-based Bayesian meta-analysis of fMRI studies, and evaluated its performance by examining the overlap index using the output from NeuroSynth, a tool for synthesis of large datasets from fMRI articles. We also compared the result from Bayesian meta-analysis with that from classical meta-analysis. The comparison demonstrated that Bayesian meta-analysis showed a higher overlap index than the classical meta-analysis did, particularly when 2logBF = 6 was used for thresholding. Given our assumption pertaining to the meaning of the overlap index, this result implies that Bayesian meta-analysis performed better when compared to the classical meta-analysis. This result is consistent with a previous neuroimaging study that examined the performance of Bayesian second-level analysis. In the study, the best performance was produced when 2logBF = 6 threshold indicating presence of strong evidence supporting existence of significant effect was used (Han & Park, 2018). Of course, classical inference reported slightly better outcomes when thresholds more stringent than $p < .05$ were applied. However, as Salimi-Khorshidi et al. (2009) mentioned, $p < .05$ with FDR correction has been regarded as a suggested threshold for end users in the case of IBMA due to the statistical power issue, and the more stringent thresholds were employed in this study to examine extreme cases for explorations. Hence, from the perspectives of end users, Bayesian IBMA would perform better compared with classical IBMA in ordinary circumstances. The improved performance of Bayesian meta-analysis might originate from the nature of Bayesian inference that is robust against fluctuations in sample sizes and variations in the data (Han & Park, 2018). Furthermore, although the small size of our dataset, six images, may limit the credibility of analysis outcomes, our Bayesian meta-analysis demonstrated that it could show a better performance compared with classical inference within such a limited circumstance.

Interestingly, when we analyzed an additional dataset with twelve images, which also includes six images created by adding the random noise to the original six images, Bayesian IMBA outperformed Classical IBMA even when the forward inference image was used for the performance evaluation. Although Classical IBMA showed a slightly better performance compared with Bayesian inference when extremely stringent thresholds were applied (i.e., $p < .01$, $.005$, or $.001$ after FDR correction) and the forward inference NeuroSynth image was used, we could not see such a result in the case of this additional performance evaluation. The results from the additional performance evaluation may suggest that first, Bayesian analysis is more robust against the possible noise existing in images compared with classical analysis when the image quality is suboptimal as argued by Han and Park (2018), and second, Bayesian IMBA can show better performance compared with classical IMBA as a sample size increases.

In addition to the improved performance, we stress that employing Bayes hypothesis testing has epistemological benefits as well. To recapitulate, a p-value is all about the null; it does not tell us whether the alternative hypothesis is true given the observed data. As discussed earlier, however, hypothesis testing via BF allow us to directly compare the null and alternative hypotheses, and to accept one of them when the BF is sufficiently small or large. In that process, the researcher has to make the alternative hypothesis explicit and clear, which is another benefit of adopting the Bayesian hypothesis testing framework. Furthermore, the possibility of accepting the null is not a feature which the classical hypothesis testing framework can afford; it is an exclusive benefit of using BFs for hypothesis testing. Considering the goal of neuroscientific research in the context of fMRI studies where one has to determine if a voxel became active or *not* under some cognitive or affective task, such a benefit must be taken seriously.

Several limitations remain to be addressed. One major issue is the limitation associated with the sample size and data availability. Availability of shared fMRI image files could be a significant issue hindering the application of image-based Bayesian

meta-analysis. Although NeuroVault has been developed as a repository for image files reporting results from statistical analyses, we could not download sufficient image files for our meta-analysis from the database. Of course, although we could show that Bayesian meta-analysis could work better than classical meta-analysis with the limitation of the small sample size and the results were not significantly altered in our prior robustness check, further analyses should be done with more datasets within more diverse functional domains for better generalization. A study using full simulation data might be a possible way to address this issue. Although we conducted a performance evaluation with an additional dataset that included simulation data containing the random noise, we could not use a dataset that solely consisted of simulation data in the present study. In fact, we were mainly interested in how Bayesian meta-analysis would work from the perspective of end users, who might attempt to analyze their own real data. In addition, we intended to utilize results from previous meta-analyses based on larger datasets as standards for our performance evaluation. Therefore, evaluating meta-analysis methods using fully hypothetical simulation data could not be a viable option because it is practically impossible to find previous meta-analyses using full simulation data whose property and nature are identical to full simulation data that we would create and meta-analyze. Given these reasons, we considered that utilizing full simulation data was out of the scope of the present study. More considerations regarding the possible application of the simulation method to evaluate the performances of meta-analysis methods may need to be done in future studies.

In addition to the aforementioned major limitation, there are minor limitations that should be considered as well. First, the development of a user-friendly tool for Bayesian meta-analysis was out of the scope of the present study and yet to be achieved. End-users may need to modify the R code shared via GitHub to perform their own meta-analysis. Second, because it is well-known that choice of prior could significantly affect the result in Bayesian analysis (Sinharay & Stern, 2002), it is necessary to carefully examine the impact

of prior on the result of meta-analysis. Of course, our prior robustness check reported that the BF outcomes were not sensitive to the change in the prior distribution; however, we only examined Cauchy and normal distributions, so other distributions may need to be tested in future research. Third, although one of the benefits of using Bayesian analysis is that it is possible to set prior distributions based on findings from previous studies, we could not do so in the present study. In fact, a previous study suggested that prior distributions could be formed from results from meta-analysis of relevant previous studies (Jones, Ades, Sutton, & Welton, 2018). However, we could not utilize the aforementioned approach in our study because, first, in the case of fMRI analysis, it is required to set a prior distribution for each individual voxel, and second, we intended to use results from previous meta-analysis as standards for performance evaluation so it would be inappropriate to use the same meta-analysis data to form prior distributions. Researchers may need to consider how to implement such an approach in Bayesian fMRI analysis in future studies.

Declaration of interest

Declarations of interest: none

References

Berger, J. et al. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. doi:10.1002/jrsm.12

Cui, X., Li, J., & Song, X. (2015). Xjview. Retrieved June 28, 2015, from http://www.alivelearn.net/xjview

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7900–7905. doi:10.1073/pnas.1602413113

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. doi:10.1080/19345747.2011.618213

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138. doi:10.1080/23743603.2017.1326760

Han, H. & Glenn, A. L. (2018). Evaluating methods of correcting for multiple comparisons implemented in SPM12 in social neuroscience fMRI studies: an example from moral psychology. *Soc. Neurosci. 13*(3), 257–267. doi:10.1080/17470919.2017.1324521

Han, H. & Park, J. (2018). Using SPM 12's second-level bayesian inference procedure for fMRI analysis: Practical guidelines for end users. *Frontiers in Neuroinformatics*, *12*. doi:10.3389/fninf.2018.00001

Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2017). metaBMA: Bayesian Model
      Averaging for Random and Fixed Effects Meta-Analysis. R package version 0.3.9.
      Retrieved from https://cran.r-project.org/package=metaBMA

Jones, H. E., Ades, A. E., Sutton, A. J., & Welton, N. J. (2018). Use of a random effects
      meta-analysis in the design and analysis of a new clinical trial. *Statistics in Medicine*.
      doi:10.1002/sim.7948

Kang, J., Johnson, T. D., Nichols, T. E., & Wager, T. D. (2011). Meta Analysis of
      Functional Neuroimaging Data via Bayesian Spatial Point Processes. *Journal of the
      American Statistical Association*, *106*(493), 124–134. doi:10.1198/jasa.2011.ap09735

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical
      Association*, *90*(430), 773–795. doi:10.2307/2291091. arXiv: 1099-1255

Lieberman, M. D. & Cunningham, W. A. (2009). Type I and Type II error concerns in
      fMRI research: Re-balancing the scale. *Social, Cognitive and Affective Neuroscience*,
      4, 423–428. doi:10.1093/scan/nsp052

Quintana, D. S. & Williams, D. R. (2018). Bayesian alternatives for common
      null-hypothesis significance tests in psychiatry: a non-technical guide using JASP.
      BMC Psychiatry, *18*(1), 178. doi:10.1186/s12888-018-1761-4

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A., Schulz, J., . . . Eickhoff, S.
      (2012). Modelling neural correlates of working memory: A coordinate-based
      meta-analysis. *NeuroImage*, *60*(1), 830–846. doi:10.1016/j.neuroimage.2011.11.050

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t
      tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*,
      16(2), 225–237.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., & Nichols, T. E. (2009).
      Meta-analysis of neuroimaging data: A comparison of image-based and
      coordinate-based pooling of studies. *Neuroimage*, *45*(3), 810–823.
      doi:10.1016/j.neuroimage.2008.12.039

Samartsidis, P., Eickhoff, C. R., Eickhoff, S. B., Wager, T. D., Barrett, L. F., Atzil, S., . . .
Nichols, T. E. (2018). Bayesian log-Gaussian Cox process regression: applications to
meta-analysis of neuroimaging working memory studies. *The Journal of the Royal
Statistical Society, Series C (Applied Statistics)*. doi:10.1111/rssc.12295

Scott, J. G. & Berger, J. O. (2006). An exploration of aspects of bayesian multiple testing.
Journal of Statistical Planning and Inference, *136*(7), 2144–2162.

Sinharay, S. & Stern, H. S. (2002). On the sensitivity of bayes factors to the prior
distributions. *The American Statistician*, *56*(3), 196–201.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package.
Journal of Statistical Software, *36*(3). doi:10.18637/jss.v036.i03

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values.
Psychonomic bulletin & review, *14*(5), 779–804.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A bayesian perspective on the
bonferroni adjustment. *Biometrika*, *84*(2), 419–427.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J.
(2011). Statistical evidence in experimental psychology: An empirical comparison
using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298.

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

Yue, Y. R., Lindquist, M. A., & Loh, J. M. (2012). Meta-analysis of functional
neuroimaging data using Bayesian nonparametric binary regression. *The Annals of
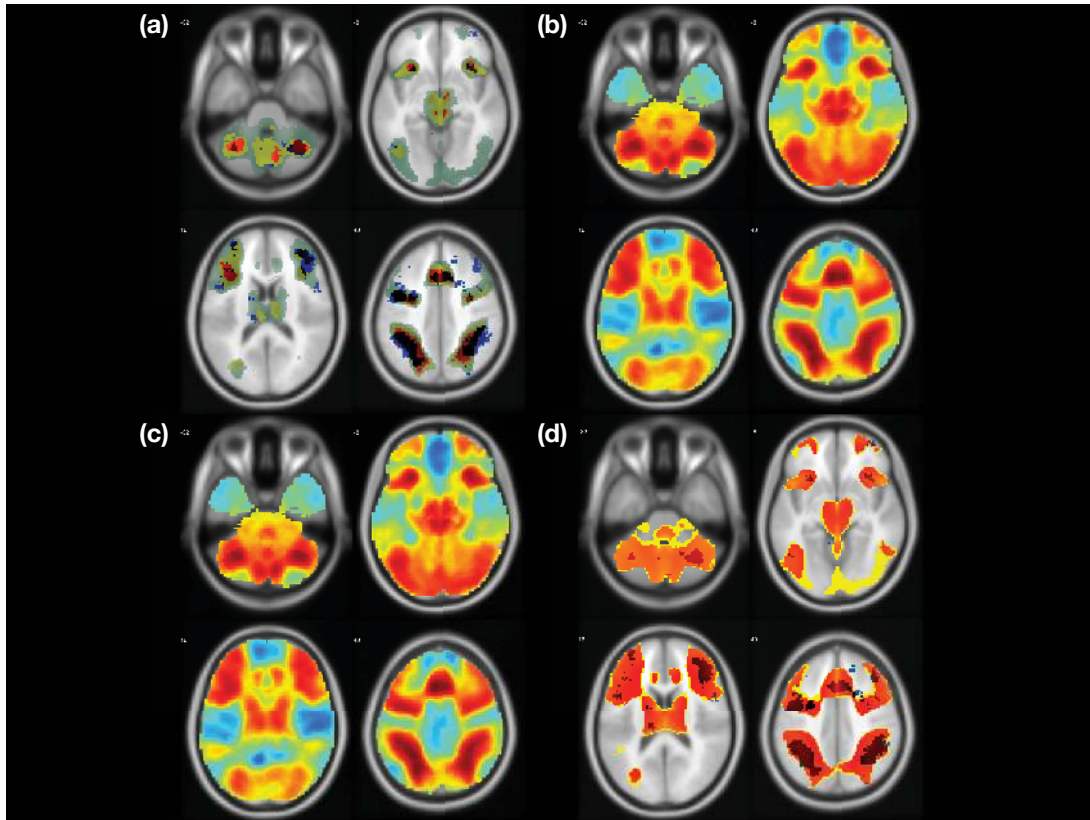Applied Statistics*, *6*(2), 697–718. doi:10.1214/11-AOAS523

Figure 1. Results from Bayesian and classical meta-analysis. (a) Thresholded mask image from Bayesian meta-analysis. Red: voxels survived 2logBF = 10. Yellow: voxels survived 2logBF = 6. Green: voxels survived 2logBF = 2. Blue to black: NeuroSynth result. (b) Mean effect size image from Bayesian meta-analysis. (c) Median effect size image from Bayesian meta-analysis. (d) Thresholded mask image from classical meta-analysis. Red to orange: voxels survived FDR .01. Yellow to orange: voxels survived FDR .05. Blue to black: NeuroSynth result. / Online-only color for all figures
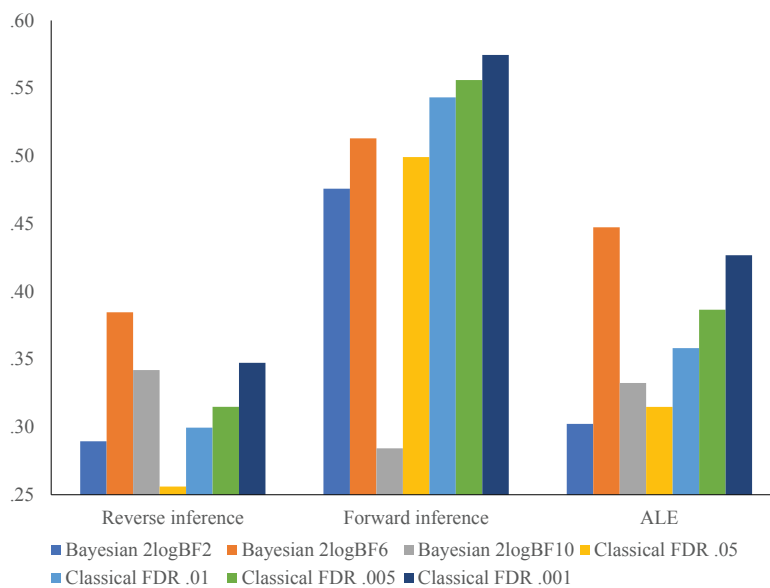
(a) Overlap index comparison between Bayesian
and classical meta-analysis with different
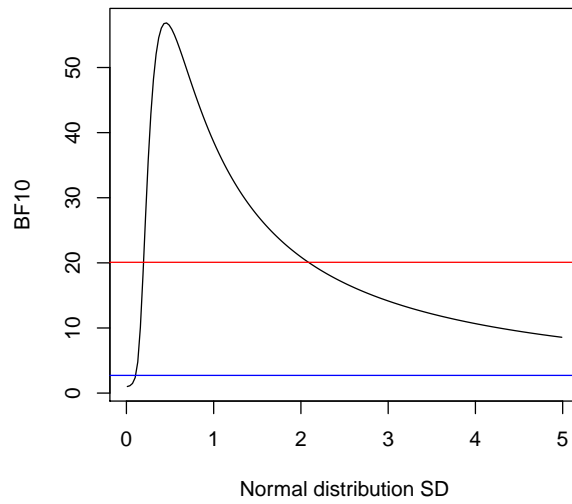thresholds.
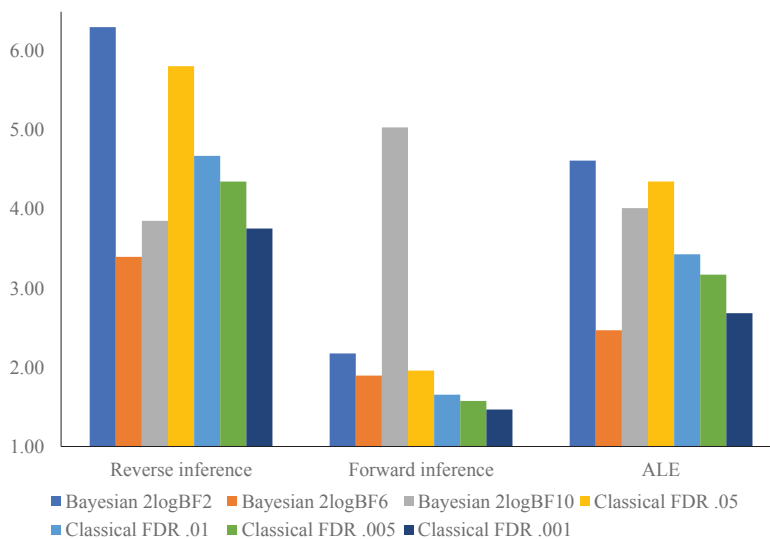
(b) Ratio of error voxels to correct voxels
comparison between Bayesian and classical
meta-analysis with different thresholds.

Figure 2. Performance evaluation results
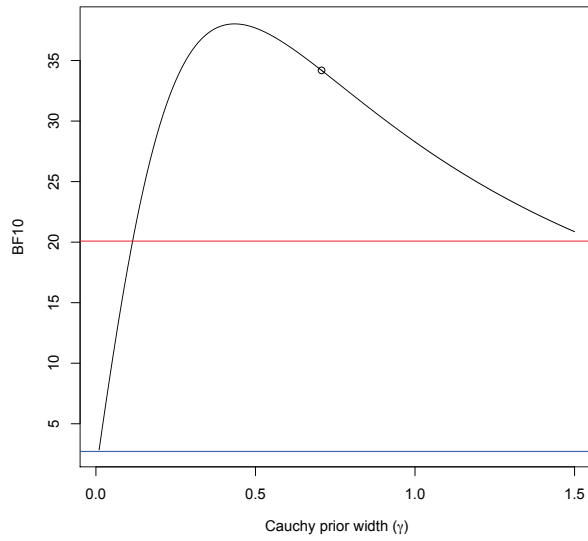
(a) Cauchy distribution prior. Circle represents
the actual resultant BF value with the default
Cauchy prior distribution.

(b) Normal distribution prior.

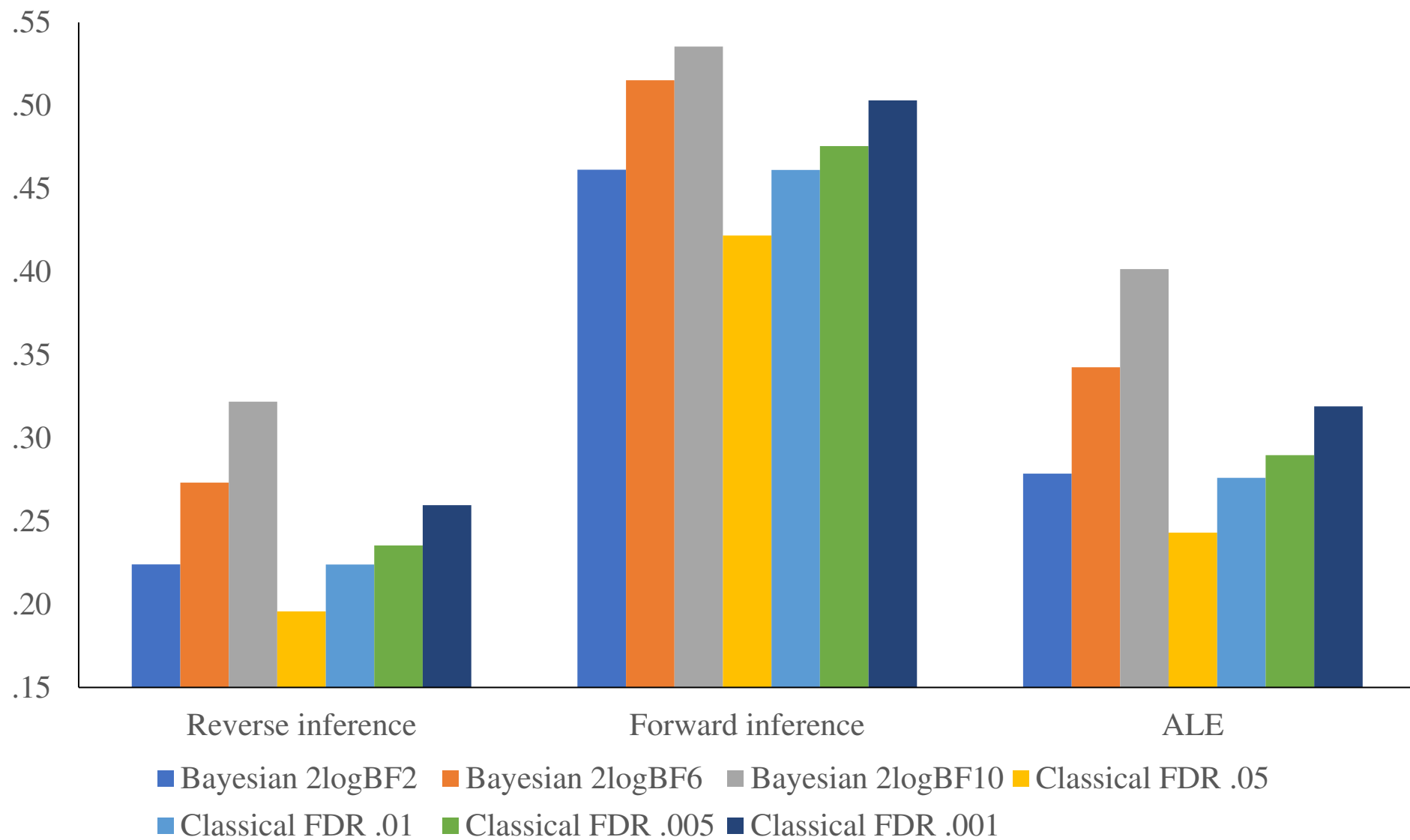*Figure 3*. Prior robustness check results

Figure S1. Overlap index comparison between Bayesian and classical meta-analysis with different threshold when noise-added simulation data was added.
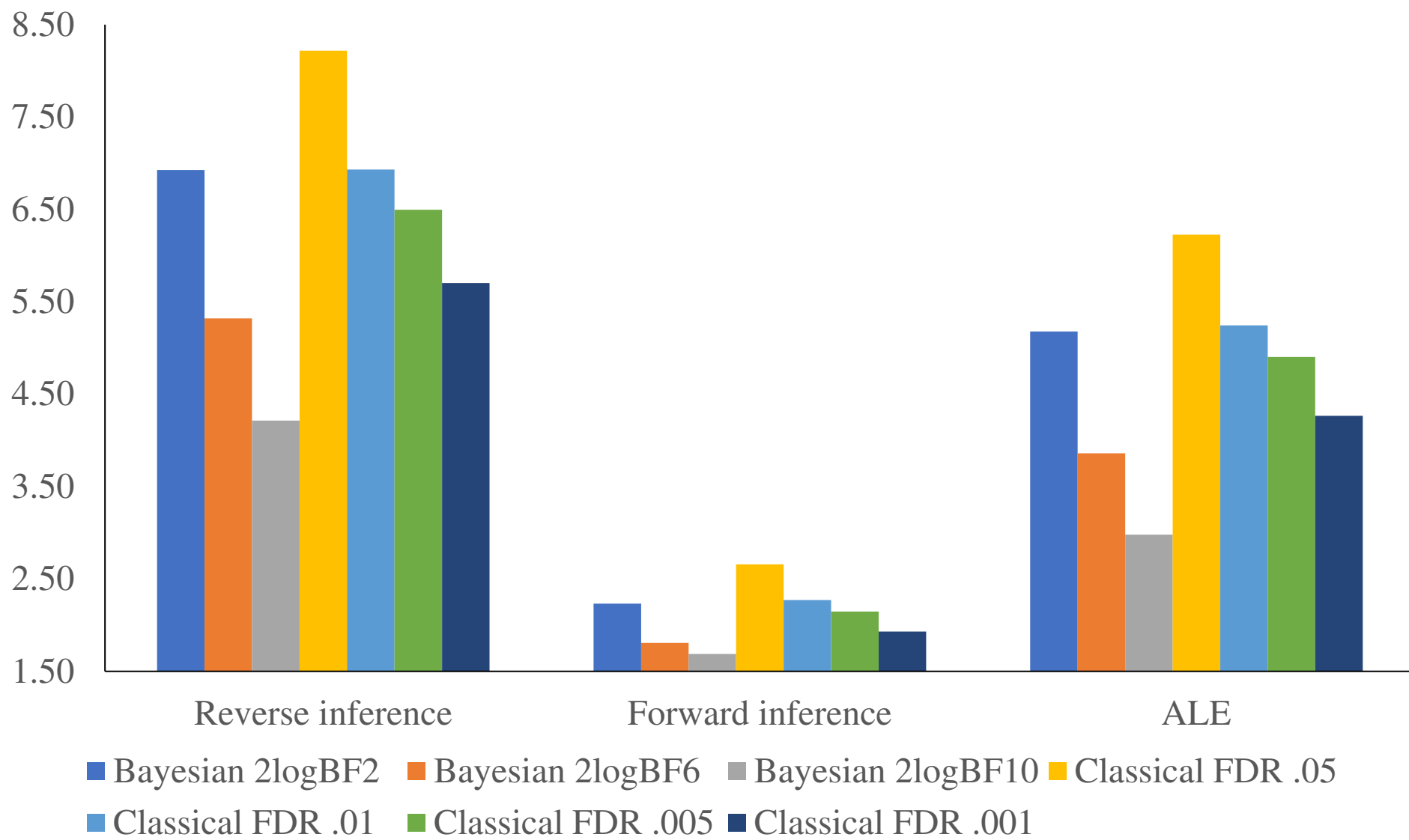
Figure S2. Ratio of error voxels to correct voxels comparison between Bayesian and classical meta-analysis with different threshold when noise-added simulation data was added.