RESEARCH

A Minimum Bayes Factor Based Threshold for Activation Likelihood Estimation

Tommaso Costa[1,2,3] · Donato Liloia[1,2] · Franco Cauda[1,2,3] · Peter T. Fox[4,5] · Francesca Dalla Mutta[2] · Sergio Duca[1,2] · Jordi Manuello[1,2]

Abstract

Activation likelihood estimation (ALE) is among the most used algorithms to perform neuroimaging meta-analysis. Since its first implementation, several thresholding procedures had been proposed, all referred to the frequentist framework, returning a rejection criterion for the null hypothesis according to the critical p-value selected. However, this is not informative in terms of probabilities of the validity of the hypotheses. Here, we describe an innovative thresholding procedure based on the concept of minimum Bayes factor (mBF). The use of the Bayesian framework allows to consider different levels of probability, each of these being equally significant. In order to simplify the translation between the common ALE practice and the proposed approach, we analised six task-fMRI/VBM datasets and determined the mBF values equivalent to the currently recommended frequentist thresholds based on Family Wise Error (FWE). Sensitivity and robustness toward spurious findings were also analyzed. Results showed that the cutoff $\log_{10}(mBF) = 5$ is equivalent to the FWE threshold, often referred as voxel-level threshold, while the cutoff $\log_{10}(mBF) = 2$ is equivalent to the cluster-level FWE (c-FWE) threshold. However, only in the latter case voxels spatially far from the blobs of effect in the c-FWE ALE map survived. Therefore, when using the Bayesian thresholding the cutoff $\log_{10}(mBF) = 5$ should be preferred. However, being in the Bayesian framework, lower values are all equally significant, while suggesting weaker level of force for that hypothesis. Hence, results obtained through less conservative thresholds can be legitimately discussed without losing statistical rigor. The proposed technique adds therefore a powerful tool to the human-brain-mapping field.

Keywords Activation likelihood estimation · Minimum Bayes Factor · BrainMap · Coordinate-based meta-analysis · GingerALE

Introduction

Since the seminal work of Ogawa et al. (1990), research based on magnetic resonance imaging (MRI) has produced a large amount of data associated to cognitive domains and sensory processes, in both healthy and diseased populations. However, the high complexity of the topic investigated, together with peculiarities of the used techniques, expose results to a potential high variability (Bowring et al., 2019; Smith et al., 2005). This issue can be overcome by means of coordinate-based meta-analysis (CBMA), a technique allowing to pool results from different experiments that investigated similar questions (Wager et al., 2007, 2009). Notably, this allows to produce additional statistics based on a much greater sample size of those analyzed in the original experiments, which sometimes leads to scarcely reliable individual results (Manuello et al., 2022).

Among the available techniques to perform CBMA, activation likelihood estimation (ALE) aims to estimate the spatial convergence of activation probabilities between the results of previously published experiments. It is based on the null-hypothesis that the foci (i.e., the peaks of effect reported in each single experiment) are uniformly spread

✉ Donato Liloia
  donato.liloia@unito.it

[1] GCS-fMRI Group, Koelliker Hospital and Department of Psychology, University of Turin, Turin, Italy

[2] FOCUS Laboratory, Department of Psychology, University of Turin, Via Verdi 10, 10124 Turin, Italy

[3] Neuroscience Institute of Turin, Turin, Italy

[4] Research Imaging Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

[5] Biggs Institute for Alzhiemer's and Neurodegenerative Diseases, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

across the brain, while looking for regions in which the convergence across multiple experiments is higher than if it came from an independent distribution (Eickhoff et al., 2012). One of the crucial elements of ALE is the thresholding procedure, and for this reason, it evolved over time to improve the statistical reliability of the results obtained by means of this meta-analytic approach. In the first implementation, the voxel-wise significance was assessed based on a fixed-effect analysis without correction for multiple comparisons. Since the lack of this correction was likely to introduce false positives, successive modifications introduced the false discovery rate (FDR) approach to account for this (Laird et al., 2005). Later on, Eickhoff et al. (2009) empirically estimated the spatial variability of data for the Gaussian filtering and optimized the implementation of the permutation test. Finally, Eickhoff et al. (2012) integrated the method with two additional thresholding procedures based on the correction for family-wise error (FWE) rate and the cluster-level significance (c-FWE). All these inferential procedures are theoretically based on the conjoint use of Fisher's p-value approach and Neyman-Pearson's Test approach, which are often combined in scientific practice.

Here, we aimed to introduce an innovative thresholding option, based on Bayesian inference rather than on the frequentist one. Specifically, this leverages on the concept of minimum Bayes Factor (mBF), which is the smallest Bayes Factor obtained for a p-value in a given distribution with respect to the alternative hypothesis. The mBF has several advantages over the available canonical thresholding procedures. On the statistical side, the canonical ALE-related thresholds, which are based on frequentist inference, solely provide a rejection criterion for the null hypothesis according to the critical p-value selected. However, this is not informative in terms of probabilities of the validity of hypotheses (Cauda et al., 2020; Costa et al., 2021; Liloia et al., 2022). On the contrary, the Bayesian solution allows to consider different levels of probability, each of these being equally significant. Moreover, the BF, which is a unique function of the data, allows a straightforward interpretation, as it expresses how probable it is to find a real effect in a given voxel. Secondarily, on the computational side, it does not require any random or permutation test. This is particularly relevant when analyzing large datasets of experiments, allowing to significantly reduce the required processing time (i.e., a few seconds versus tens of minutes, up to hours in worst cases).

In the present work, we first provided a formal description of the BF and the mBF. Then, aiming to facilitate the comparison between the common ALE practice and the proposed new procedure, we analyzed a range of meta-analytic datasets to identify the mBF equivalent to canonical thresholds. Based on these values, we compared the sensitivity of the mBF with that of other available solutions. Finally, we tested the robustness of the Bayesian approach toward false positives detection. The introduction of this innovative thresholding procedure could therefore further consolidate one of the most used methodologies in CBMA research, allowing a deeper interpretation of the results, while simplifying at the same time its computational implementation.

Materials and Methods

Anatomical Likelihood Estimation and Creation of the Modeled Activation Maps

An in-depth description of the ALE methodology (Eickhoff et al., 2009, 2012; Turkeltaub et al., 2012) is out of scope for the present work. Here, we briefly recapitulate its core concepts following the nomenclature used in Samartsidis et al. (2017). In the ALE environment, a Gaussian distribution is centered on each voxel, representing the probability to find a true effect in it. In other words, this means computing the likelihood of the voxel to be the exact location of a focus reported in one (or more) of the original experiments included in the dataset. Given the experiment i, the map based on one of its focus $\boldsymbol{x}_{ik}$ is expressed as:

$$L_{ik}(v) = c\phi_3(v|\boldsymbol{x}_{ik}, \sigma_i^2\boldsymbol{I}) \tag{1}$$

where k is a focus in the experiment considered; $\phi_3(\boldsymbol{x};\mu,\boldsymbol{\Sigma})$ is a three-dimensional Gaussian distribution with mean and covariance μ, Σ evaluated in $\boldsymbol{x} \in \mathbb{R}^3$. c is a normalization constant ensuring that the sum of ϕ_3 over voxels equals to 1, and I is the identity matrix. Therefore, $L_{ik}(v)$ represents the likelihood that the voxel v is the true location of the focus $\boldsymbol{x}_{ik}$. The only free parameter in formula (1) is σ_i. Its value was empirically determined in Eickhoff et al. (2009), considering the number of subjects n_i in each experiment.

After this first step, the L_{ik} maps obtained for each single focus reported in a given experiment are combined in L_i, also named modelled activation (MA) map. Each MA represents the estimate of the full effect map obtained in the original experiment and described in a reduced form through the associated list of foci. Finally, the ALE statistic is computed as follows:

$$l(v) = 1 - \prod_i (1 - L_i(v)) \tag{2}$$

This formula represents the probability for the voxel v to be the location of a true effect coherently measured across the experiments included in the dataset. This algorithm is implemented in GingerALE software (https://www.brainmap.org/ale/) (Eickhoff et al., 2009, 2012; Turkeltaub et al., 2012). In addition to the unthresholded and thresholded ALE maps, the output also includes a Z map. The conversion from p-values to Z points is based on the numerical method proposed by Cody (1969).

Bayes Factor and the Minimum Bayes Factor

The computation of the mBF is based on the Bayes' theorem. According to it, the probability associated with two concurrent hypotheses, namely H_0 and H_1, can be formalized as follows:

$$P(H_0|D) = \frac{P(D|H_0)}{P(D)}P(H_0) \tag{3}$$

and, correspondingly,

$$P(H_1|D) = \frac{P(D|H_1)}{P(D)}P(H_1) \tag{4}$$

where D is the measured effect in a voxel.

Their quotient represents the Bayes' theorem in terms of relative belief:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}\frac{P(H_0)}{P(H_1)} \tag{5}$$

Based on the previous formulas, the BF_{01} can therefore be expressed as:

$$BF_{01} = \frac{P(D|H_0)}{P(D|H_1)} \tag{6}$$

The value of BF_{01} gives the degree of evidence for the two concurrent hypotheses: when $BF_{01} > 1$ the evidence favors H_0; on the contrary, when $BF_{01} < 1$ the evidence favors H_1.

In the usual form, using the Bayes Factor to measure the evidence for a composite hypothesis requires averaging all possible distinct alternative values. However, in some cases infinite alternative values exist, as for instance for the hypothesis "the mean is different from zero". Therefore, determining a point value among all the possible ones of the composite hypothesis can be extremely complex and sometime impractical. To circumvent this issue, it is possible to select the alternative value with the largest effect against the null hypothesis, considering this as the summary of evidence across all the possible distinct values contained in the composite alternative hypothesis. This is called the Minimum Bayes' Factor (mBF), which therefore reflects the largest amount of evidence against the null hypothesis. Of note, Eq. (6) is in the simple form of a likelihood ratio. Therefore, the mBF uses the best-supported alternative hypothesis H_1. This represents the worst-case scenario for H_0 among all possible distinct Bayes' factors values, because no alternative value is supported by a larger amount of evidence against H_0 than the one identified through the mBF.

The BF and the mBF are usually based on the data D, as notated in Eqs. (3–6). However, it is equally possible to use the mBF on test statistics or p-values. Therefore, Eq.

(6) still holds true replacing the data D with p-values, or these can be back-transformed to the underlying statistics t or z and then proceeding with the calculation. This transformation is one-to-one; $z = z(p)$ is well defined and it can be seen how the mBF is unaffected by using either z or p.

Let's now consider as the hypothesis the probability of a generic observed effect x modelled through a Gaussian distribution with mean μ and variance σ^2:

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

the BF for the null hypothesis versus the supported hypothesis ($\mu = x$) is the mBF:

$$\frac{P(x|\mu=0,\sigma)}{P(x|\mu=x,\sigma)} = \frac{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-0)^2}{2\sigma^2}}}{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-x)^2}{2\sigma^2}}} = e^{-\frac{x^2}{2\sigma^2}} \tag{8}$$

If we now take the Z map obtained from the unthresholded ALE map, we can compute the mBF applying the following formula:

$$mBF_{01} = e^{-\frac{Z^2}{2}} \tag{9}$$

by applying a simple exponentiation. This represents the evidence of the null hypothesis H_0 against the alternative hypothesis H_1. To obtain the evidence of H_1 the reciprocal of the mBF_{01} can be computed as:

$$mBF_{10} = \frac{1}{mBF_{01}} \tag{10}$$

The mBF ranges from 0 to ∞. According to Kass and Raftery (1995) values can be interpreted as shown in Table 1. The MATLAB® code which implements the Z to mBF transformation described in formula 10 is freely available on figshare. (https://figshare.com/articles/software/minimum_bayes_factor_script/17023931).

Identification of the Equivalent mBF Thresholding

Since its inception, the ALE has been based on frequentist statistics, and researchers applying it are therefore used to reason in frequentist terms. Therefore, in order to facilitate the mental translation between the p-value based interpretation and Bayesian one, we first aimed to find mBF values equivalent to canonical thresholds applied to ALE. To do so, the ALE algorithm was applied to six different datasets, three of them referring to task-related functional MRI (fMRI) data, the remaining three including only voxel-based morphometry (VBM) experiments. In order to maximize reliability and reproducibility, five of these datasets were

Table 1 The evidence categories for the Bayes Factor (BF). Adapted from Kass and Raftery (1995)

BF_{10} value	Force of evidence
1—3	Very weak
3—20	Positive
20—150	Strong
>150	Very strong

selected from published CBMAs based on ALE method. Briefly, the first task-fMRI pool consisted of 109 experiments on time-perception, originally collected and analyzed in Nani et al. (2019). A second functional pool included 73 experiments based on finger-tapping task, from Laird et al. (2008). The third and last one gathered 19 experiments on face discrimination, from Eickhoff et al. (2012). On the structural side, the largest pool included 57 experiments on Alzheimer's disease, previously analyzed in Manuello et al. (2018). A further pool focused on chronic schizophrenia, with 49 experiments collected in Liloia et al. (2021). Finally, the last pool was built querying the VBM section of the BrainMap database (Vanasse et al., 2018) to retrieve 20 available experiments on mild cognitive impairment (MCI) (see also Table S1 in the Supplement). It should be noted, that the selected datasets had varying sizes. This was decided in order to verify the possible confounding effect of the number of experiments analysed. In particular, the face discrimination and MCI datasets were selected to verify the behavior of mBF threshold when the sample size approaches the lower bound recommended for valid ALE analyses (Eickhoff et al., 2016).

The identification of the equivalent mBF threshold was then based on a four steps procedure, separately applied to each dataset: i) Data were analyzed through GingerALE (v.3.0.2), applying different canonical thresholds with the default settings suggested by the tool. To be more precise, for the fixed-effect uncorrected thresholding the p-value was set to 0.05. For the cluster-level thresholding the uncorrected p-value was set to 0.001, the corrected p-value to 0.05, tested with 1000 permutations. For the voxel-level thresholding the corrected p-value was set to 0.05, tested with 1000 permutations. ii) The thresholded ALE maps obtained in the previous step were converted into binary masks. iii) The Z map computed by GingerALE was converted into a BF map, as shown in formulas 9–10. iv) The BF map was individually masked with each of the previously binarized thresholded ALE maps. At this point, the lowest voxel value in the obtained output was taken as the mBF threshold. In other words, this is the cutoff that can be set to apply the Bayesian equivalent of the canonical threshold. In terms of interpretation, the BF values in the map express how much more probable it is to find a true convergence of effect in a given voxel as compared to finding no convergence. In order to prove its reliability, the mBF equivalent to a given canonical threshold must be consistent across the different datasets analysed.

Correlation between BF and Canonical ALE Maps

For the purpose of further investigating the behavior of the Bayesian approach, we aimed to identify the mBF value that maximizes the correlation with the ALE maps canonically thresholded. In details, ALE maps were computed, thresholded and binarized as described above in steps i and ii. The BF map was obtained as in step iii, and then expressed in logarithmic scale. This $\log_{10}(BF)$ map was iteratively binarized applying increasing threshold values, and Pearson correlation was run with the binarized ALE map at each step. Notably, the mBF previously obtained matches the mBF returning the maximum correlation only if the mBF-thresholded map maximizes the overlap with the canonical ALE map, while minimizing non-overlapping voxels.

Sensitivity of the Equivalent mBF Thresholding

As previously described, the mBF threshold was identified by means of masking procedure base on the canonical ALE maps. However, this does not exclude in a non-masked map the presence of further voxels with the same mBF value but non-overlapping with the canonical ALE map of reference. Therefore, in order to be a suitable alternative to the frequentist thresholds, the mBF must produce maps covering the highest possible fraction of the canonical ALE map, while minimizing non-overlapping voxels. Moreover, the potentially remained non-overlapping voxels should be localized close to the clusters of overlapping voxels, rather than randomly spread throughout the brain. These can be qualitatively assessed by a visual comparison of the mBF and canonical ALE maps. To obtain a quantitative measurement, the $\log_{10}(mBF)$ map was iteratively binarized at increasing threshold values as described above, and at each step the number of voxels appearing only in the mBF map, only in the ALE map, or in both of them was counted. Finally, the $\log_{10}(mBF)$ threshold allowing to remove any voxel appearing in the Bayesian map but not in the canonical ALE map was computed.

Robustness Towards Spurious Convergences

Since the ALE method aims to identify spatial convergence among multiple experiments, the frequentist thresholds were designed to suppress possible spurious convergences in the results. As the same behavior is required from the Bayesian approach, we tested it on a dataset including 20 experiments that investigated unrelated cognitive domains (see Table S2 for a description of this dataset), and on two simulated datasets generated with the Fail-safe R script (Acar et al. (2018); https://github.com/NeuroStat/GenerateNull) and built with 21 experiments each with foci randomly placed across the brain. In this condition, possible voxels surviving the threshold must be considered as spurious evidence, therefore the mBF approach is expected to produce empty maps.

Results

Identification of the Equivalent mBF

With the aim of identifying an equivalent mBF value for the different canonical thresholds, six datasets were processed with ALE algorithm and thresholded with both Bayesian and canonical frequentist approaches. Results showed a consistent behavior of the mBF (Table 2). Specifically, the p05 threshold was shown to be equivalent to a mBF of ≈4. As shown in Table 1, this is rather weak evidence, consistent with the already known low reliability of its equivalent frequentist solution. Since the use of p05 thresholding is to date explicitly discouraged in the standard practice of ALE (Eickhoff et al., 2016), no further analyses were performed on for its Bayesian counterpart. The c-FWE threshold was shown to be equivalent to a mBF of ≈119. Finally, the FWE threshold was shown to be equivalent to a mBF of ≈10^5. Of note, the equivalent mBF value for c-FWE and FWE can be interpreted as 'strong' and 'very strong' evidence, respectively. No differences were found between task-fMRI and structural datasets, and no effect of sample size was detected. The iterative correlation between the maps thresholded using either the Bayesian approach or the canonical one confirmed the previously determined mBF values. In most cases, peaks exceeded r = 0.9, considering either FWE or c-FWE (Fig. 1). Interestingly, in the two datasets with sample size around the recommended minimum (i.e., MCI and face discrimination), the correlation peaks with FWE were considerably higher than those obtained with c-FWE.

Sensitivity and Robustness

The qualitative comparison between the canonical ALE map and the equivalent mBF-thresholded map showed, in line with the observed correlation peaks, a substantial overlap between the two. However, the spatial distribution of the non-overlapping mBF voxels differed between FWE and c-FWE. In the former case, the non-overlapping voxels surviving the threshold of $\log_{10}(mBF) = 5$ were localized around the regions of overlap (Fig. S1). In contrast, voxels surviving the threshold of $\log_{10}(mBF) = 2$ in the latter case formed blobs far from the regions of overlap (Fig. S2). Notably, in order to suppress voxels that were present in the mBF map but not in the c-FWE ALE map, the $\log_{10}(mBF)$ threshold had to be increased by orders of magnitude, being at least 4 for finger tapping and 8 for time perception (Table 3; Fig. 2). However, this also removed up to the 94.5% of the voxels in overlap, in the worst case. This loss of convergence was much lower in the case of FWE ALE maps, never exceeding the 59%. Finally, it is relevant to note that in all cases the equivalent mBF threshold marks the end of the steady state of the number of overlapping voxels between the mBF map and the canonical ALE map. This same landmark also coincides with the correlation peak between the two conditions.

The analyses of the random tasks dataset and of the two random foci datasets confirmed that a $\log_{10}(mBF) = 2$ is not conservative enough to suppress spurious convergences. On the contrary, no voxels in the maps reached a value of $\log_{10}(mBF) = 5$ (See Fig. S3).

Discussion

In the present work, we described and applied an innovative thresholding approach for ALE meta-analyses based on the mBF. The adoption of the Bayesian framework is a relevant development for the CBMAs field, which had included so far only the frequentist paradigm. In fact, although these two statistical families often tend to produce convergent findings (Jaynes & Kempthorne, 1976), they differ in the logic used to interrogate data and interpret results (Cauda et al., 2020; Costa et al., 2021; Liloia et al., 2022; Poldrack, 2006). For this reason, while suggesting a new framework we first aimed to identify its quantitative equivalent in the frequentist counterpart. This consistently emerged across the

Table 2 mBF values equivalent to canonical frequentist thresholds, obtained as described in the paragraph "Identification of the equivalent mBF thresholding". In the last two columns, the maximum Pearson correlation value obtained, as described in the paragraph "Correlation between BF and canonical ALE maps", between the Bayesian and canonical ALE maps is shown (this information is also graphically represented in Fig. 1). Values in brackets refer to the number of experiments in the dataset

MRI Modality	Dataset	equivalent minimum_BF			max correlation	
		p05	FWE	c-FWE	FWE	c-FWE
Task fMRI	Time perception (109)	3.8711	1.97E+05	1.19E+02	0.95	0.95
	Finger tapping (73)	3.8816	2.12E+05	1.19E+02	0.96	0.98
	Face discrimination (19)	3.8891	1.13E+05	1.19E+02	0.99	0.90
VBM	Alzheimer (57)	3.8724	2.35E+05	1.19E+02	0.94	0.92
	Schizophrenia (49)	3.8754	2.44E+05	1.19E+02	0.86	0.83
	MCI (20)	3.872	9.96E+04	1.19E+02	0.98	0.85

Fig. 1 The iterative Pearson correlation between the maps thresholded with either the Bayesian approach or the canonical one. The blue line refers to the correlation with the FWE thresholding, while the orange dashed line refers to the correlation with the c-FWE thresholding. The x axis represents the $\log_{10}$(mBF) value used to threshold the Bayesian map. Values in brackets refer to the number of experiments in the dataset. MCI = mild cognitive impairment
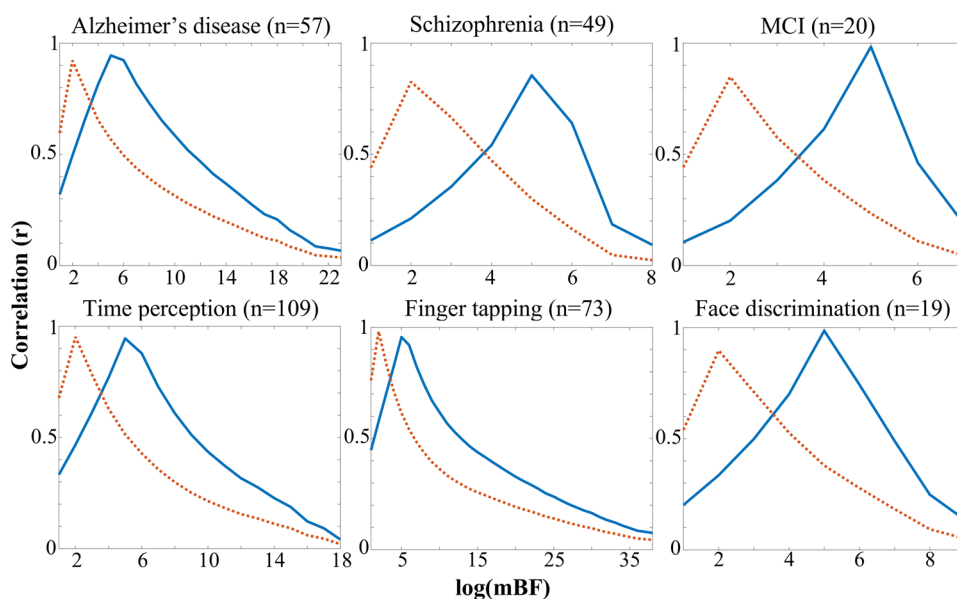
six heterogeneous meta-analytic datasets analysed, showing the mBF≈4 cutoff being tantamount to the p05 canonical threshold, the mBF≈119 cutoff being tantamount to the c-FWE canonical threshold and the mBF≈10^5 cutoff being tantamount to the FWE canonical threshold.

Although the interpretation of BF values proposed by Kass and Raftery (1995) was conceived for a general scope and is therefore not specific for neuroimaging results, it is possible to note that the force of the evidence retrievable through the equivalent of the p05 thresholding is "very weak". On the contrary, results obtained with mBF≈119

Table 3 $\log_{10}$(mBF) values necessary to suppress voxels present in the Bayesian map but not in the canonical ALE map, together with the percentage of overlapping voxel lost while increasing the Bayesian threshold from the equivalent mBF previously identified to the more stringent suppressing threshold. The corresponding procedure was explained in the paragraph "Sensitivity of the equivalent mBF thresholding". Values in brackets refer to the number of experiments in the dataset

MRI Modality	Dataset	FWE		c-FWE	
		non overlap suppression th	% of overlap lost	non overlap suppression th	% of overlap lost
Task fMRI	time_perception (109)	6	22.8	8	91.0
	finger_tapping (73)	6	15.2	4	49.0
	face_discrimination (19)	6	44.7	5	85.6
VBM	Alzheimer (57)	6	14.7	6	75.3
	Schizophrenia (49)	6	59.0	5	91.0
	MCI (20)	5	3.3	5	94.5

(i.e., c-FWE) are scored as "strong", and those obtained with mBF≈10^5 (i.e., FWE) are "very strong". Notably, Kass and Raftery's cutoff for "very strong" evidence is BF > 150, while values in our results are at least three orders of magnitude greater. This information, that was not obtainable through the frequentist approach, reveals that the ALE technique can produce extremely robust results, in the Bayesian meaning. Moreover, it seems to suggest the canonical FWE thresholding being more conservative than the c-FWE option. It is relevant to note that when ALE is applied to unrelated or random data, results did not reach the same strength of evidence observed for real datasets.

Since one of the advantages of ALE is to be adequate for both functional and structural MRI data, describing both the healthy or the pathological brain (Eickhoff et al., 2012) the consistent behavior of the mBF thresholding observed across all the tested datasets was strongly necessary. Similarly, the mBF approach showed to be reliable for the whole range of sample sizes being valid for ALE analysis (i.e., at least 17 experiments per dataset as recommended by Eickhoff et al. (2016)). However, contrary to the frequentist framework, the Bayesian approach doesn't require a minimum number of evidences to produce unbiased results (Costa et al., 2021). For this reason, provided that every meta-analysis aims to include the largest possible fraction of the published literature on a given topic, the mBF thresholding now allows to perform CBMA in those specific domains which had been so far hampered by an insufficient number of retrievable experiments.

As a further advantage, the final ALE maps obtained through the mBF approach are easily comparable among them, as the values directly express the strength of the evidence. For example, the observed BF value of 200 in

Fig. 2 Number of non-zero voxels appearing only in the mBF-thresholded map (blue line), only in the canonically-thresholded map (yellow line), or in overlap between the two maps (red line). The x axis represents the $\log_{10}$(mBF) value used to threshold the Bayesian map. The vertical dashed line marks the $\log_{10}$(mBF) threshold equivalent to canonical thresholds as previously identified. Values in brackets refer to the number of experiments in the dataset. MCI = mild cognitive impairment
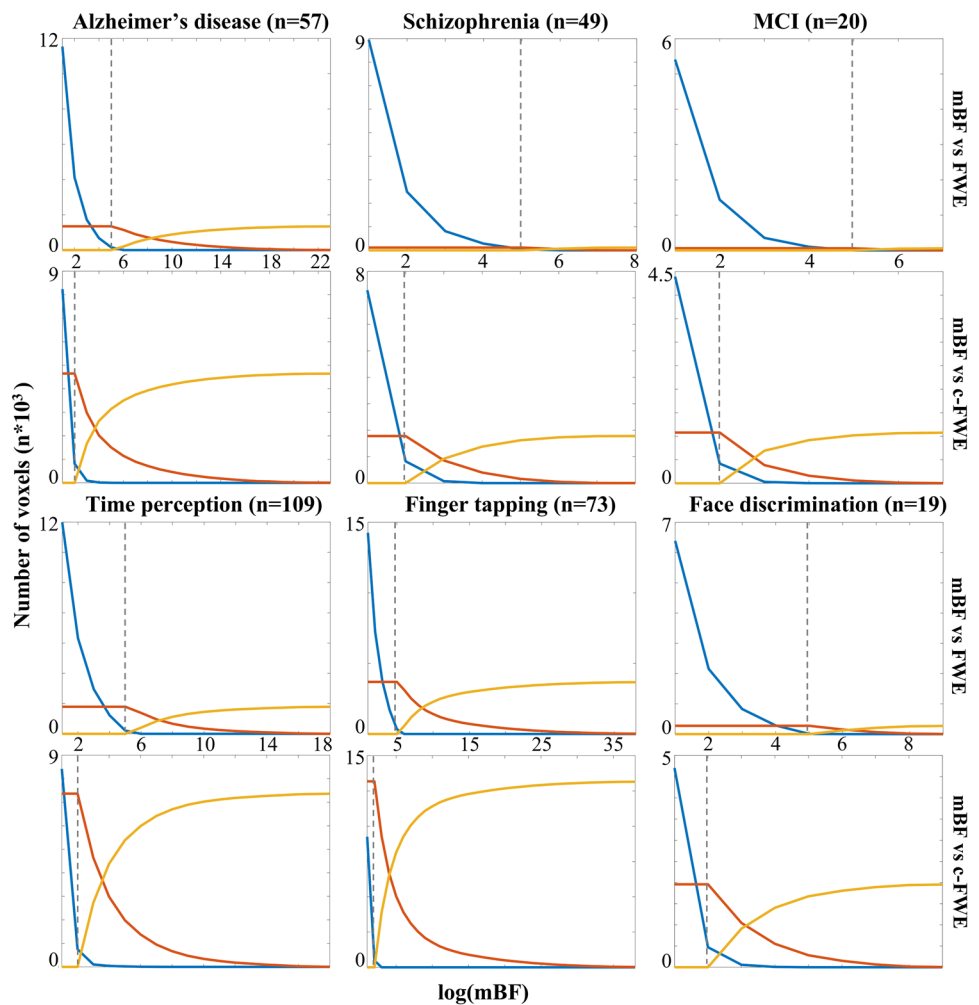
a voxel activated or altered by the 'condition A' would be exactly twice as strong of the BF value of 100 observed in a voxel activated or altered by the 'condition B'. Extending the same linear relationship observed in the Bayesian case to the frequentist paradigm would be technically wrong. In fact, although sequentially developed over time, the p05, the FWE, and the c-FWE are based on different statistics (Eickhoff et al., 2016) and results obtained with them should not be interpreted as being lying on a continuum, despite the discussed trend of equivalent mBF values.

Similarly, the logic behind lowering the mBF threshold is not equivalent to the meaning of modifying the p-value. In fact, evidence obtained in the frequentist framework at $p < 0.01$ is 100 times less accurate than evidence obtained at $p < 0.0001$. On the contrary, evidence showing in the Bayesian framework a value of mBF = 1 is 100 times weaker than evidence showing a value of mBF = 100, but the two are equally accurate. The same difference should be considered when testing a-priori hypotheses. Let's suppose, as an example, to expect to find the involvement of brain region A in a given cognitive domain, but then to

not observe that in the final ALE map thresholded at FWE $p < 0.001$. Increasing the p-value to 0.01 or even to 0.05 until the expected brain region appears in the map is a questionable procedure, that should be clearly highlighted by researchers. On the contrary, if that region doesn't survive mBF = 10^5 it is licit to lower the threshold to whichever value, then discussing the observed strength for the hypothesis of its involvement in that cognitive domain. In light of this, there is no absolute cutoff to be recommended for the use of the mBF based ALE thresholding. However, the following standards should be followed to report and describe results obtained through mBF thresholding in a clear and transparent way: i) No binarization should be applied to the thresholded maps; ii) No upper-bound cutoff should be used to limit the range of values for visualization purpose (logarithmic scale can be used instead, if needed); iii) The maps should be always accompanied with a color scale clearly showing the range of values. If it is true that the mBF threshold can be adjusted, it is extremely relevant to know how the mBF values are distributed across the map. The implications of the recommended standards
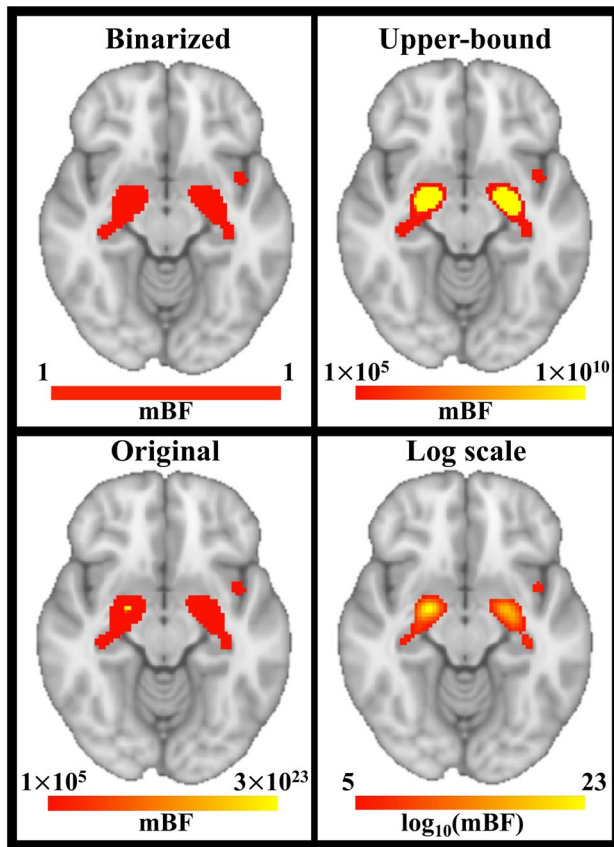
Fig. 3 A visual representation of the recommended standards to report results in the Bayesian framework. In each condition the map was thresholded at mBF = 10^5. In the top left quadrant, the mBF = 10^5 threshold was also used to binarize the map. In the top right quadrant, an arbitrary upper-bound cutoff at mBF = $10.^{10}$ was applied. The bottom left quadrant shows the original map. In the bottom right quadrant, the logarithmic scale was used instead. Results refers to the analysed dataset on Alzheimer's disease. The axial slice is shown in radiological standard (left is right)

can be appreciated in Fig. 3. Applying the binarization, the small cluster in the left insula seems to be equally informative as the other two cluster, and it is impossible to localize any peak in the map. Similarly, the peaks remain hidden if setting an upper-bound cutoff. Conversely, the original map, arbitrarily thresholded at mBF = 10^5, allows to identify the peak of effect, although the rest of the maps appears flattened around the threshold. On the contrary, resorting to the logarithmic scale allows to appreciate the full distribution of values.

As previously mentioned, and following Kass and Raftery (1995), any evidence showing BF > 150 should be considered as "very strong". At the same time, our results showed voxels with value greater than $\log_{10}(mBF) = 35$, being hardly comparable with Kass and Raftery's range of 1–150. Therefore, as the standard practice in the ALE field has been so far based on frequentist thresholds, it sounds reasonable to highlight that the use of a mBF threshold of 10^5 warranties results strongly comparable with those obtained through the FWE thresholding, generally considered a reliable standard (Eickhoff et al., 2016). On the contrary, a mBF threshold of 10^2 shouldn't be intended as tantamount to the c-FWE thresholding. This is likely ascribable to the lack of a topological criteria in the Bayesian approach, which is instead applied in the c-FWE algorithm after the initial uncorrected p-value estimate (Eickhoff et al., 2012). Finally, if compared with the computation of FWE-based thresholds in the context of ALE, mBF has the advantage of not being dependent on Monte-Carlo method, which is computationally intense and therefore potentially resulting in highly time-consuming analyses. While proposing an alternative procedure for statistical thresholding, the mBF approach builds on the ALE technique, and therefore inherits its limitations, as well as those of CBMAs in general. The main one pertains to the creation of the dataset, which should be as extensive as possible. Nevertheless, the Bayesian approach is less prone to the effects of potentially biased sampling, as it explicitly considers the specific set of data, without referring to any hypothetical whole population. As suggested by our results, the proposed approach should be equally efficient irrespective of the content of the domain analysed. While 6 datasets could be deemed as insufficient amount of data to provide reliable evidence in terms of generalizability, it should be noted that their selection was carefully designed to maximize their representativeness. We therefore included data describing both cognitive domains and pathologies, through functional or structural MRI data respectively. Moreover, the cognitive domains and related tasks were selected to be highly heterogeneous, and both psychiatric and neurodegenerative disorders were covered. As an additional point, the datasets were gathered between 2008 and 2021, therefore potentially capturing the effects of technical changes happening during the years. Finally, we also manipulated the number of experiments included in each dataset, which is commonly considered the factor to have more power to affect and bias results obtained through ALE. It is of course not possible to exclude a discrepancy of the equivalent mBF for canonical thresholds in specific datasets, but since there are no technical or theoretical reasons to expect markedly divergent behavior to happen systematically, there was no way to set an ideal number of datasets to be analysed. Although the shared code to implement the transformation from Z maps to BF maps was written in MATLAB®, the same procedure can be adapted to any other language and programming environment. Finally, while we referred to the GingerALE software, any Z map produced as output of other implementations of the ALE algorithm is equally suitable for the mBF approach here described.

Conclusions

CBMA is a widely used and meaningful technique in human neuroimaging. The adoption of a Bayesian framework for statistical thresholding, and specifically of the computation of mBF, can strengthen the ALE approach to CBMAs, expanding its field of application to small datasets, and further confirming the robustness of results obtained through this technique.

Information Sharing Statement

The data that support the findings of this study can be freely downloaded from https://figshare.com/articles/dataset/dataset_minimum_bayes_factor_zip/21621162.The code used to perform the analyses can be freely downloaded from https://figshare.com/articles/software/minimum_bayes_factor_script/17023931.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Acar, F., Seurinck, R., Eickhoff, S. B., & Moerkerke, B. (2018). Assessing robustness against potential publication bias in activation likelihood estimation (ALE) meta-analyses for fMRI. *PLoS ONE, 13*(11), e0208177. https://doi.org/10.1371/journal.pone.0208177

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping, 40*(11), 3362–3384. https://doi.org/10.1002/hbm.24603

Cauda, F., Nani, A., Liloia, D., Manuello, J., Premi, E., Duca, S., & Costa, T. (2020). Finding specificity in structural brain alterations through Bayesian reverse inference. *Hum Brain Mapp, n/a*(n/a). https://doi.org/10.1002/hbm.25105

Cody, W. J. (1969). Rational Chebyshev approximations for the error function. *Mathematics of Computation, 23*(107), 631–637.

Costa, T., Manuello, J., Ferraro, M., Liloia, D., Nani, A., Fox, P. T., & Cauda, F. (2021). BACON: A tool for reverse inference in brain activation and alteration. *Human Brain Mapping, 42*(11), 3343–3351. https://doi.org/10.1002/hbm.25452

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage, 59*(3), 2349–2361. https://doi.org/10.1016/j.neuroimage.2011.09.017

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping, 30*(9), 2907–2926. https://doi.org/10.1002/hbm.20718

Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., & Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage, 137*, 70–85. https://doi.org/10.1016/j.neuroimage.2016.04.072

Jaynes, E. T., & Kempthorne, O. (1976). Confidence Intervals vs Bayesian Intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: Proceedings of an International Research Colloquium held at the University of Western Ontario, London, Canada, 10–13 May 1973 Volume II Foundations and Philosophy of Statistical Inference* (pp. 175–257). Dordrecht: Springer Netherlands.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., & Fox, P. T. (2005). ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum Brain Mapp, 25*(1), 155–164. https://doi.org/10.1002/hbm.20136

Laird, A. R., Robbins, J. M., Li, K., Price, L. R., Cykowski, M. D., Narayana, S., & Fox, P. T. (2008). Modeling motor connectivity using TMS/PET and structural equation modeling. *Neuroimage, 41*(2), 424–436. https://doi.org/10.1016/j.neuroimage.2008.01.065

Liloia, D., Brasso, C., Cauda, F., Mancuso, L., Nani, A., Manuello, J., & Rocca, P. (2021). Updating and characterizing neuroanatomical markers in high-risk subjects, recently diagnosed and chronic patients with schizophrenia: A revised coordinate-based meta-analysis. *Neuroscience & Biobehavioral Reviews, 123*, 83–103. https://doi.org/10.1016/j.neubiorev.2021.01.010

Liloia, D., Cauda, F., Uddin, L. Q., Manuello, J., Mancuso, L., Keller, R., & Costa, T. (2022) Revealing the Selectivity of Neuroanatomical Alteration in Autism Spectrum Disorder via ReverseInference. *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/j.bpsc.2022.01.007

Manuello, J., Nani, A., Premi, E., Borroni, B., Costa, T., Tatu, K., & Cauda, F. (2018). The pathoconnectivity profile of Alzheimer's Disease: A morphometric coalteration network analysis. *Front Neurol, 8*(739). https://doi.org/10.3389/fneur.2017.00739

Manuello, J., Costa, T., Cauda, F., & Liloia, D. (2022) Six actions to improve detection of critical features for neuroimaging coordinate-based meta-analysis preparation. *Neuroscience & Biobehavioral Reviews*, 104659. https://doi.org/10.1016/j.neubiorev.2022.104659

Nani, A., Manuello, J., Liloia, D., Duca, S., Costa, T., & Cauda, F. (2019). The Neural Correlates of Time: A Meta-analysis of

Neuroimaging Studies. *Journal of Cognitive Neuroscience, 31*(12), 1796–1826. https://doi.org/10.1162/jocn_a_01459

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences, 87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Samartsidis, P., Montagna, S., Nichols, T. E., & Johnson, T. D. (2017). The coordinate-based meta-analysis of neuroimaging data. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 32*(4), 580. https://doi.org/10.1214/17-STS624

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., & McGonigle, D. J. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human brain mapping, 24*(3), 248–257.https://doi.org/10.1002/hbm.20080

Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., & Fox, P. (2012). Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses. *Human Brain Mapping, 33*(1), 1–13. https://doi.org/10.1002/hbm.21186

Vanasse, T. J., Fox, P. M., Barron, D. S., Robertson, M., Eickhoff, S. B., Lancaster, J. L., & Fox, P. T. (2018). BrainMap VBM: An environment for structural meta-analysis. *Human Brain Mapping, 39*(8), 3308–3325. https://doi.org/10.1002/hbm.24078

Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: Current and future directions. *Social Cognitive and Affective Neuroscience, 2*(2), 150–158. https://doi.org/10.1093/scan/nsm015

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & Van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage, 45*(1), S210–S221. https://doi.org/10.1016/j.neuroimage.2008.10.061