JNeurosci THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Power-up: a reanalysis of 'power failure' in neuroscience using mixture modelling

Camilla L Nord[1], Vincent Valton[1], John Wood[2] and Jonathan P Roiser[1]

[1]*Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, UK, WC1N 3AZ*
[2]*Research Department of Primary Care and Population Health, University College London Medical School, Rowland Hill Street, London, UK, NW3 2PF*

DOI: 10.1523/JNEUROSCI.3592-16.2017

Received: 22 November 2016

Revised: 6 June 2017

Accepted: 17 June 2017

Published: 13 July 2017

Author contributions: C.L.N., J.W., and J.R. designed research; C.L.N. and V.V. performed research; C.L.N. and V.V. analyzed data; C.L.N., V.V., and J.R. wrote the paper.

Conflict of Interest: JPR is a consultant for Cambridge Cognition and Takeda. CLN, VV, and JW declare no competing interests.

The authors would like to thank Karl Friston, Geraint Rees, James Kilner, and Oliver Robinson for comments on an earlier draft of the manuscript, Katherine Button and Marcus Munafò for their invaluable help with the replication portion of the analysis, and Oon Him Peh for assistance with publication bias analyses. The authors also gratefully acknowledge the reviewers of the manuscript for their helpful comments.

Correspondence should be addressed to Please address correspondence to: Camilla L Nord, 17 Queen Square, London WC1N 3AZ, camilla.nord.11@ucl.ac.uk, +442076791138

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.3592-16.2017

Power-up: a reanalysis of 'power failure' in neuroscience using mixture modelling

Abbreviated title: Power-up

Nord, Camilla L[1*], Valton, Vincent[1*], Wood, John[2], Roiser, Jonathan P[1]

[1]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, UK, WC1N 3AZ

[2] Research Department of Primary Care and Population Health, University College London Medical School, Rowland Hill Street, London, UK, NW3 2PF

*These authors contributed equally to this work.

Please address correspondence to:

Camilla L Nord

17 Queen Square, London WC1N 3AZ

camilla.nord.11@ucl.ac.uk

+442076791138

No. pages: 31

No. figures: 5; no. tables: 2

No. words abstract: 169

Conflict of interest: JPR is a consultant for Cambridge Cognition and Takeda. CLN, VV, and JW declare no competing interests.

Acknowledgements: The authors would like to thank Karl Friston, Geraint Rees, James Kilner, and Oliver Robinson for comments on an earlier draft of the manuscript, Katherine Button and Marcus Munafò for their invaluable help with the replication portion of the analysis, and Oon Him Peh for assistance with publication bias analyses. The authors also gratefully acknowledge the reviewers of the manuscript for their helpful comments.

Abstract

Evidence for endemically low statistical power has recently cast neuroscience findings into doubt. If low statistical power plagues neuroscience, this reduces confidence in reported effects. However, if statistical power is not uniformly low, such blanket mistrust might not be warranted. Here, we provide a different perspective on this issue, analysing data from an influential paper reporting a median power of 21% across 49 meta-analyses (Button et al., 2013). We demonstrate, using Gaussian mixture modelling, that the sample of 730 studies included in that analysis comprises several subcomponents; therefore the use of a single summary statistic is insufficient to characterise the nature of the distribution. We find that statistical power is extremely low for studies included in meta-analyses that reported a null result; and that it varies substantially across subfields of neuroscience, with particularly low power in candidate gene association studies. Thus, while power in neuroscience remains a critical issue, the notion that studies are systematically underpowered is not the full story: low power is far from a universal problem.

Significance statement

Recently, researchers across the biomedical and psychological sciences have become concerned with the reliability of results. One marker for reliability is statistical power: the probability of finding a statistically significant result, given that the effect exists. Previous evidence suggests that statistical power is low across the field of neuroscience. Our results present a more comprehensive picture of statistical power in neuroscience: on average, studies are indeed underpowered—some very seriously so—but many studies show acceptable or even exemplary statistical power. We show that this heterogeneity in statistical power is common across most subfields in neuroscience (psychology, neuroimaging, etc.). This new, more nuanced picture of statistical power in neuroscience could affect not only scientific understanding, but potentially policy and funding decisions for neuroscience research.

Introduction

Trust in empirical findings is of vital importance to scientific advancement, but publishing biases and questionable research practices can cause unreliable results (Nosek et al., 2012; Button et al., 2013). In recent years, scientists and funders across the biomedical and psychological sciences have become concerned with what has been termed a crisis of replication and reliability (Barch and Yarkoni, 2013).

One putative marker for the reliability of results is statistical power: the probability that a statistically significant result will be declared, given that the null hypothesis is false (i.e., a real effect exists). It can be shown that, in the context of field-wide underpowered studies, a smaller proportion of significant findings will reflect true positives than if power is universally high (Ioannidis, 2005). A recent influential paper by Button and colleagues (Button et al., 2013) calculated statistical power across all meta-analyses published in 2011 that were labelled as “neuroscience” by Thomson Reuters Web of Science. It concluded that neuroscience studies were systematically underpowered, with a median statistical power of 21%, and that the proportion of statistically significant results that reflect true positives is therefore likely to be low. The prevalence of very low power has serious implications for the field. If the majority of studies are indeed underpowered, statistically significant findings are untrustworthy, and scientific inference will often be misinformed. This analysis provoked considerable debate in the field about whether neuroscience does indeed suffer from endemic low statistical power (Bacchetti, 2013; Quinlan, 2013). We sought to add nuance to this debate by re-analysing the original dataset using a more fine-grained approach, and provide a different perspective on statistical power in neuroscience.

We extended the analyses of Button and colleagues (Button et al., 2013), using data from all 730 individual studies, which provided initial results that were consistent with the original report (which used only the median-sized study in 49 meta-analyses). To quantify the heterogeneity of the dataset we made use of Gaussian mixture modelling (GMM) (Corduneanu and Bishop, 2001), which assumes that the data may be described as being composed of multiple Gaussian components. We then used model comparison to find the most parsimonious model for the data. We also categorised each study based on its methodology to examine whether low power is common to all fields of neuroscience.

We find strong evidence that the distribution of power across studies is multi-modal, with the most parsimonious model tested including four components. Moreover, we show that candidate gene association studies and studies from meta-analyses with null results make up the majority of extremely low powered studies in the analysis of Button and colleagues. Although median power in neuroscience is low, the distribution of power is heterogeneous, and there are clusters of adequately and even well-powered studies in the field. Thus, our in-depth analysis reveals that the crisis of power is not uniform: instead, statistical power is extremely diverse across neuroscience.

Methods

Experimental design and analysis

Re-analysing 'power failures'

Our initial analysis took a similar approach to that of Button and colleagues, but contrary to their protocol (which reported power only for the median-sized study in each meta-analysis: N=49), we report power for each of the 730 individual studies (see Figure 3a and Table 1). As in the original analysis, we defined power as the probability that a given study would declare a significant result, assuming that the population effect size was equal to the weighted mean effect size derived from the corresponding meta-analysis (note that this differs from 'post-hoc' power, in which the effect size would be assumed to be equal to the reported effect size from each individual study (O'Keefe, 2007)).

For experiments with a binary outcome, power was calculated by assuming that the expected incidence or response rate for the control group (i.e. the base rate) was equal to that reported in the corresponding meta-analysis and, similarly, used an assumed "treatment effect" (odds or risk ratio) equal to that given by each meta-analysis. The test statistic used for the calculation was the log odds-ratio divided by its standard error. The latter was derived from a first order approximation, and estimated by the square root of the sum of the reciprocals of the expected values of the counts in the 2-by-2 summary table. The test statistic itself was then referenced to the standard normal distribution for the purposes of the power calculation. For studies reporting Cohen's *d*, the assumed treatment effect was again taken directly from the corresponding meta-analysis, and all power calculations were based on the standard noncentral *t*-distribution. For comparability with the original study we calculated the median power across all 730 individual studies which was equal to 23%, close to the 21% reported by Button and colleagues (2013).

Figure 1 shows an overview of our analytical process. We additionally classified each study according to methodology: candidate gene association studies (N=234); psychology (N=198); neuroimaging (N=65); treatment trials (N=145); neurochemistry (N=50); and a miscellaneous category (N=38 studies from N=2 meta-analyses). Two independent raters categorized the 49 meta-analyses into these six subfields, with 47/49 classified consistently; the remaining two were resolved following discussion. Before continuing our analysis in more depth, we present the reader with results that are directly comparable with the analysis of Button and colleagues (with the addition of the subfields; Table 2). These results are intended for comparison with our more nuanced characterisation of the distributions using GMMs presented below; given the results of those GMMs (which suggest the these distributions are multi-modal and therefore not well characterised by a single measure of central tendency) they should not be used to draw strong inferences.

Figure 1. Classification of studies for analysis

Description of study methodology. GMM=Gaussian mixture model.

First author of study	*k*	Cohen's *d*	Odds ratio	CI	Significance	Classification
Babbage (Babbage et al., 2011)	13	-1.11		-0.97 to -1.25	*	Psychology
Bai (Bai, 2011)	18		1.47	1.22 to 1.77	*	Genetic
Bjorkhelm-Bergman (Björkhem-	6	-1.20		1.6 to 8.0	*	Treatment
Bucossi (Bucossi et al., 2011)	21	.41		.17 to .65	*	Neurochemistry
Chamberlain (Chamberlain et	11	-.51		.825 to 1.08	*	Psychology
Chang (Chang et al., 2011a)	56	-.19		-.29 to -.1	*	Psychology
Chang (Chang et al., 2011b)	6		.98	.86 to 1.12	-	Genetic
Chen (Chen et al., 2011)	12		.6	.52 to .69	*	Miscellaneous
Chung (Chung and Chua, 2011)	11		.67	.43 to 1.04	-	Treatment
Domellof (Domellöf et al., 2011)	14		2.12	1.59 to 2.78	*	Psychology
Etminan (Etminan et al., 2011)	14		0.8	.7 to .92	*	Treatment
Feng (Feng et al., 2011)	4		1.20	1.04 to 1.4	*	Genetic
Green (Green et al., 2011)	17	-.59		-.93 to -.257	*	Neurochemistry
Han (Han et al., 2011)	14		1.35	1.06 to 1.72	*	Genetic
Hannestad (Hannestad et al.,	13	-.13		-.55 to .29	-	Treatment
Hua (Hua et al., 2011)	27		1.13	1.05 to 1.21	*	Genetic
Lindson (Lindson and Aveyard,	8		1.05	.92 to 1.19	-	Treatment
Liu (Liu et al., 2011a)	12		1.04	.88 to 1.22	-	Genetic
Liu (Liu et al., 2011b)	6		.89	.82 to .96	*	Genetic
MacKillop (MacKillop et al.,	57	.58		.509 to .641	*	Psychology
Maneeton (Maneeton et al.,	5		1.67†	1.23 to 2.26	*	Treatment
Ohi (Ohi et al., 2011)	6		1.12	1.00 to 1.26	*	Genetic
Olabi (Olabi et al., 2011)	14	-.4		-.62 to -.19	*	Brain imaging
Oldershaw (Oldershaw et al.,	10	-.51		-.73 to -.28	*	Psychology
Oliver (Oliver et al., 2011)	7		.86	0.79 to .95	*	Treatment
Peerbooms (Peerbooms et al.,	36		1.26	1.09 to 1.46	*	Genetic
Pizzagalli (Pizzagalli, 2011)	22	.92		.442 to 1.393	*	Treatment
Rist (Rist et al., 2011)	5		2.06	1.33 to 3.19	*	Miscellaneous
Sexton (Sexton et al., 2011)	8	.43		.063 to .799	*	Brain imaging
Shum (Shum et al., 2011)	11	.89		.75 to 1.02	*	Psychology
Sim (Sim et al., 2011)	2		1.23†	1.08 to 1.52	*	Treatment
Song (Song et al., 2011)	12	.15		.043 to .264	*	Neurochemistry
Sun (Sun et al., 2011)	6		1.93	1.55 to 2.41	*	Genetic
Tian (Tian et al., 2011)	4	1.26		.947 to 1.568	*	Treatment
Trzesniak (Trzesniak et al.,	11		1.98	1.33 to 2.94	*	Brain imaging
Veehof (Veehof et al., 2011)	8	.37		.20 to .53	*	Treatment
Vergouwen (Vergouwen et al.,	24		.83	.74 to .93	*	Treatment
Vieta (Vieta et al., 2011)	10		.68†	.60 to .77	*	Treatment
Wisdom (Wisdom et al., 2011)	53	-.14		-.21 to -.07	*	Genetic
Witteman (Witteman et al.,	26	-1.41		-1.76 to -1.05	*	Psychology
Woon (Woon and Hedges,	24	-.60		-.83 to -.37	*	Brain imaging
Xuan (Xuan et al., 2011)	20		1.00	.861 to 1.156	-	Genetic
Yang (cohort) (Yang et al.,	14		1.38†	1.18 to 1.61	*	Miscellaneous
Yang (case control) (Yang et al.,	7		2.48	1.93 to 3.19	*	Miscellaneous
Yang (Yang et al., 2011b)	3	0.67		.43 to .92	*	Treatment
Yuan (Yuan et al., 2011)	14		4.98	3.97 to 6.23	*	Genetic
Zafar (Zafar et al., 2011)	8		1.07 †	.91 to 1.27	-	Treatment
Zhang (Zhang et al., 2011)	12		1.27	1.01 to 1.59	*	Genetic
Zhu (Zhu et al., 2011)	8	0.84		.18 to 1.49	*	Brain imaging

Table 1. Characteristics and classification of included meta-analyses

Classification performed by two independent raters. *k*: number of studies; † indicates relative risk; CI: confidence interval; * indicates p<0.05.

Group of studies	Median power (%)	Min. power (%)	Max. power (%)	2.5th and 97.5th percentile (based on raw data)	95% HDI (based on GMMs)	Total *k*
All studies	23	0.05	1	[0.05 to 1.00]	[0.00 to 0.72], [0.8 to 1.00]	730
All studies excluding null	30	0.05	1	[0.05 to 1.00]	[0.01 to 0.73], [0.79 to 1.00]	638
Genetic	11	0.05	1	[0.05 to 0.94]	[0.00 to 0.44], [0.63 to 0.93]	234
Treatment	20	0.05	1	[0.05 to 1.00]	[0.00 to 0.65], [0.91 to 1.00]	145
Psychology	50	0.07	1	[0.07 to 1.00]	[0.02 to 0.24], [0.28 to 1.00]	198
Imaging	32	0.11	1	[0.11 to 1.00]	[0.03 to 0.54], [0.71 to 1.00]	65
Neurochemistry	47	0.07	1	[0.07 to 1.00]	[0.02 to 0.79], [0.92 to 1.00]	50
Miscellaneous	57	0.11	1	[0.11 to 1.00]	[0.09 to 1.00]	38

Table 2. Median power by study type

Median, maximum, and minimum power subdivided by study type. We also provide the 2.5th and 97.5th percentile of the frequency distribution of power estimates of individual studies for the raw data and 95% highest-density intervals (95% HDI) for the GMMs. We used highest density intervals (HDI) to summarise the intervals of the most probable values from the distribution. HDIs differ from CIs in that they represent the most probable values of the distribution rather than symmetric credible intervals in a central tendency. As a result, HDIs are more suitable for summarising skewed and multimodal distributions than CIs. HDIs were computed using the HDRCDE R toolbox, which finds the shortest intervals such that these intervals encompass the 95% most probable values of the distribution. Multiple intervals may be identified if a region between modes of the distribution is unrepresentative of the distribution (i.e. below the 5% threshold) (Wand et al., 1991; Hyndman, 1996; Samworth and Wand, 2010), which occurs for multimodal data.

One or many populations?

The common measures of central tendency (mean, median, and mode) may not always characterise populations accurately, because distributions can be complex, and made up of multiple ‘hidden’ subpopulations. Consider the distribution of height in the United States: the mean is 168.8±13.04 cm (Fryar et al., 2012). This statistic is rarely reported because the distribution comprises two distinct populations: male (175.9 ±15.03 cm) and female (162.1 cm ±10.8 cm). The mean of the male population is greater than the 95th percentile of the female population. Thus, a single measure of central tendency fails to describe this distribution adequately.

In an analogous fashion, the original paper of Button and colleagues reported a median of 21% power, which could be interpreted as implying a degree of statistical homogeneity across neuroscience. The use of the median as a summary statistic, while having the straightforward interpretation of 'half above and half below', also implies that the power statistics are drawn from a distribution with a single central tendency. As we show below, this assumption is contradicted by our analyses, which makes the median statistic difficult to interpret. It should be noted that Button and colleagues themselves described their results as demonstrating a 'clear bimodal distribution'. Therefore we next explored the possibility that the power data originated from a combination of multiple distributions, using GMM.

GMM (similar to latent class analysis and factor models (Lubke and Muthén, 2005)) can be used to represent complex density functions where the central limit theorem does not apply, such as in the case of bimodal or multi-modal distributions. We fit GMMs with varying numbers of 'K' unknown components to the data and performed model selection using the Bayesian Information Criteria (BIC) scores to compare models with different fit and complexity (the higher the number of 'K' unknown components the more complex the model). This allowed us to take a data-driven approach, as opposed to direct mixture models using a set number of components: thus, we were agnostic as to the number of components that emerged from the model. The GMM with the lowest BIC identifies the most parsimonious model, trading model fit against model complexity. A difference in BIC between models of 10 or above on a natural logarithm scale is indicative of strong evidence in support of the model with the lower score (Kass and Raftery, 1995). To ensure that we used the most suitable GMM for this dataset, we ran different GMM models: standard GMMs, regularized GMMs, and Dirichlet Process GMMs (see below for full methods, and Figure 2 for model comparison, and model selection). The results were similar using each of these techniques (see Figure 2).

Finite Gaussian mixture model

For a finite GMM, the corresponding likelihood function is given by (Corduneanu and Bishop, 2001):

$$P(D|\pi,\theta) = \prod_{n=1}^{N} \left[\sum_{i=1}^{K} \pi_i \, \mathcal{N}(x_n|\theta_i) \right]$$

where π_i denotes the mixing coefficient (proportions of the *i*–th component), $\mathcal{N}(x_n|\theta_i)$ denotes the conditional probability of the observation x_n given by a Gaussian distribution with parameters θ_i and D denotes the whole dataset of observations, x_n. Generally speaking, this means that we believe that there is an underlying generative structure to the observed data, and that a mixture of Gaussian components would a reasonable description/approximation of the true generative process of this data. That is, we assume that the data D has been generated from a mixture of Gaussians distributions with varying means, variances, and weights (model parameters), which we want to uncover. To do so, we perform model inversion and find the point estimates of the model parameters that maximize the likelihood (see eq. 1 above) of the observed data (maximum likelihood estimation).

Model inversion is performed using the iterative EM (expectation-maximisation) algorithm, which finds a local maximum of the likelihood function given initial starting parameters. We performed 50 restarts with kmeans++ initialization (Arthur and

Vassilvitskii, 2007). Multiple restarts were performed in order to find the global maximum of the likelihood (i.e., the best GMM for the data; that is, the parameters that maximize the chance of observing the data), as opposed to a local maximum. This allowed us to ensure that convergence was achieved for all GMMs, on all datasets.

Traditionally, finite mixture modelling approaches require the number of components to be specified in advance of analysing the data. That is, for each finite Gaussian mixture model fitted to the data, one is required to input the number of components K present in the mixture (model inversion only estimates the parameters for each component). Finding the number of components present in the data is a model selection problem, and requires fitting multiple GMMs with varying numbers of components to the data, then comparing the model evidence for each fit, and selecting the most parsimonious model for the data in question (Bishop, 2006; Gershman and Blei, 2012; Murphy, 2012).

It is worth noting, however, that GMMs can be subject to instabilities, such as singularities of the likelihood function. Specifically, it is possible for one component to 'collapse' all of its variance onto a single data point, leading to an infinite likelihood (Bishop, 2006; Murphy, 2012) and to incorrect parameter estimation for the model. Multiple techniques have been developed in order to address this problem. The simplest and most commonly used technique is to introduce a regularization parameter. Another is to adopt a fully Bayesian approach and apply soft constraints on the possible range of likely parameter values, therefore preventing problematic and unrealistic parameter values. Both methodologies were used in this study, and we report on the resulting analysis for both implementations in the model selection section (below).

Finite Gaussian mixture model with regularization

In typical finite mixture models, a regularization parameter can be added in order to avoid likelihood singularities. To do so, a very small value is added to the diagonal of the covariance matrix, enforcing positive-definite covariance and preventing infinitely small precision parameters for individual components. This model specification enables one to address the issue of ‘collapsing’ components but also enforces simpler explanations of the data, favouring models with fewer components. The larger the regularization parameter, the simpler the models will be, as single components will tend to encompass a larger subspace of the data partition. In this study we introduced a regularization parameter of 0.001, which represents a reasonable trade-off between preventing over-fitting components to noise in the dataset, while capturing the most salient features from the data (the separate peaks); therefore providing a better generative model of the data than using non-regularized GMMs. We used this approach for our primary inferences.

Dirichlet Process Gaussian mixture model (DPGMM)

Dirichlet Process (DP) Gaussian mixture models (DPGMMs) are a class of Bayesian non-parametric methods that avoid the issue of model selection when identifying the optimal number of components in a mixture model (Gershman and Blei, 2012; Murphy, 2012). With DPGMM, we expand the original GMM model to incorporate a prior over the mixing distribution, and a prior over the component parameters (mean and variance of components). Common choices for DPGMM priors are conjugate priors such as the normal-inverse-Wishart distribution over the mean and covariance matrix of components, and a non-parametric prior over mixing proportions based on the DP.

The DP, often referred to as the Chinese restaurant process or the stick-breaking process, is a distribution over infinite partitions of integers (Gershman and Blei, 2012; Murphy, 2012). As a result, the DPGMM theoretically allows for an infinite number of components as it lets the number of components grow as the amount of data increases. The DP assigns each observation to a cluster with a probability that is proportional to the number of observations already assigned to that cluster. That is, the process will tend to cluster data points together, dependent on the population of the existing cluster and a concentration parameter α. The smaller the α parameter, the more likely it is that an observation will be assigned to an existing cluster with probability proportional to the number of elements already assigned to this cluster. This phenomenon is often referred to as the 'rich get richer'. This hyperparameter α indirectly controls how many clusters one expects to see from the data (another approach is to treat α as unknown, using a gamma hyperprior over α, and letting the Bayesian machinery infer the value (Blei and Jordan, 2006)).

Implementation and analysis for the non-regularized finite GMMs, regularized finite GMMs, and DPGMMs was performed using Matlab R2015b (Mathworks Inc.), using the Statistics and Machine Learning toolbox, the Lightspeed toolbox and the vdpgm toolbox (Kurihara et al., 2007).

Model selection

The traditional mixture modelling approach requires the number of clusters or components to be specified in advance of analysing the data. However, in many settings, including here, one does not know the number of underlying components and would like to estimate this directly from the data. One approach typically used with finite mixture models is to fit the data with varying number of components and then to select the model that provides the best trade-off between model fit (how well the model explains the data) and model complexity (how many component parameters are used in the model). A metric commonly used in this setting is the Bayesian Information Criterion (BIC), which allows one to compute an approximation to the Bayes factor (relative evidence) for a model. The BIC typically has two terms, the likelihood (how well the model fits the data) and a complexity term that penalizes more complex models with more free parameters (e.g. the number of components). The model with the lowest BIC metric is usually preferred as it provides the most parsimonious and generalizable model of the data.

For each one of the following datasets model fits were performed using non-regularized and regularized finite mixtures with up to 15 components (up to 10 components for the subfield categories – Figure 2): the original dataset; the original dataset excluding null studies; each methodological subfield within the original dataset (Genetics, Psychology, Neurochemistry, Treatment, Imaging, and Miscellaneous studies); and the original dataset excluding each methodological subfield. Model selection was then performed using the BIC in order to select the most parsimonious model for each dataset. Figure 2 presents (for each dataset) the corresponding BIC metric for increasing levels of model complexity. Plain blue lines denote the BIC metric using non-regularized GMMs, while plain red lines denote the BIC using regularized GMMs. The BIC metric curve for non-regularized GMMs (blue line) exhibits wide jumps (Figure 2), while the function should remain relatively smooth as seen with regularized-GMMs (red line). This suggests that non-regularized GMMs results were prone to overfitting and were inadequate for some of our datasets.

Finally, we compared different modelling methodologies, in order to select and report the most robust findings in terms of the estimation of the number of components. We compared non-regularized GMMs, regularized GMMs and DPGMMs on the same datasets (Figure 2), and found that regularized GMMs provided the most conservative estimation of the number of components. We therefore opted to report these results as the main findings.

Figure 2. Model comparison and model selection analysis for Gaussian mixture models (GMM), regularized GMMs and Dirichlet process GMMs (DPGMMs). The blue and red lines display Bayesian Information Criterion (BIC) scores (natural log scale) for non-regularized GMMs and regularized GMMs, respectively, for different levels of model complexity (number of mixture components). The lowest BIC score indicates the model that provides the best compromise between model fit (likelihood) and model complexity for the given dataset. Winning models for GMMs (purple dotted-dash vertical line), regularized GMMs (yellow dashed vertical line), and DPGMMs (green dotted vertical line) are clearly present for each dataset, enabling direct comparison of the output for each methodology. The regularized GMM approach provided the most parsimonious interpretation of the data on the two main datasets: all studies (a), excluding null studies (b) as well as 5 out of 6 subfield datasets – (c) to (h).

Results

We analysed the original sample of 730 powers (see histogram in Figure 3a). If the median were the most appropriate metric to describe the distribution of powers across studies, we would expect the GMM to produce a solution containing only a single component. Instead, the most parsimonious GMM solution included four components, with strong evidence in favour of this model versus either of the next best models (i.e. GMMs with 3 or 5 components - see Figure 2). Importantly, this model revealed that the overall distribution of power appears to be composed of sub-groups of lower and higher powered studies (overlay in Figure 3a). We next explored possible sources of this variability, considering the influence of both null effects and specific subfields of neuroscience.

Figure 3. Power of studies

Figure 3a-b: Histograms depicting the distribution of study powers across all 730 studies (a) and across studies excluding null meta-analyses (b). However, we note that excluding power statistics from studies included in null meta-analyses may provide an overestimation of power, because in many instances there remains uncertainty as to whether or not a true effect exists. Pale overlay: results of the regularised Gaussian mixture model (GMM), identifying four components (C1, C2, C3, C4) and their relative weights within the dataset. Below the histogram, pie charts depict methodological subfields, as well as null meta-analyses, contributing to each component. The null studies (white pie-chart sections) comprise 52 genetic studies and 40 treatment studies. The dark blue line shows the sum of the components (overall GMM prediction). c-h: histograms depicting the distribution of study powers across all meta-analyses, separated by subfield: candidate gene association studies (c); psychology studies (d); neurochemistry studies (e); treatment studies (f); imaging studies (g); miscellaneous studies (h). Pale overlays show the results of the regularised GMM for each subfield; the dark lines show the sum of the components (overall GMM prediction).

When is an effect not an effect?

The first important source of variability we considered relates to the concept of power itself. The calculation of power depends not just on the precision of the experiment (heavily influenced by the sample size), but also on the true population effect size. Logically, power analysis requires that an effect (the difference between population distributions) actually exists. Conducting a power analysis when no effect exists violates this predicate, and will therefore yield an uninterpretable result. Indeed, when no effect exists the power statistic becomes independent of the sample size and is simply equal to the Type I error rate; which by definition is the probability of declaring a significant result under the null hypothesis.

To illustrate this point, consider the meta-analysis titled ‘No association between APOE epsilon 4 allele and multiple sclerosis susceptibility’ (Xuan et al., 2011), which included a total of 5,472 cases and 4,727 controls. The median effect size (odds ratio) reported was precisely 1.00, with a 95% confidence interval from 0.861-1.156. Button and colleagues calculated the median power to be 5%, which is equal to the Type I error rate. However, as is evident from the paper's title, this meta-analysis was clearly interpreted by its authors as indicating a null effect, which is consistent with the observed result. Indeed, in this case the power is 5% for both the largest (N>3000) and the smallest (N<150) study in the meta-analysis. In such cases the estimate of 5% power is not easily interpretable.

On the other hand, it is problematic to assume that a non-significant meta-analytic finding can be taken as evidence there is no true effect; in the Frequentist statistical framework, failure to reject the null hypothesis cannot be interpreted as unambiguous evidence that no effect exists (due to the potential for false negative results). For example, reference 16 ('Effects on prolongation of Bazett's corrected QT interval of seven second-generation antipsychotics in the treatment of schizophrenia: a meta-analysis') reported a median effect size (odds ratio) of 0.67, with a 95% confidence interval from 0.43-1.04. While this result was non-significant, the point estimate of the effect size is greater than those from several meta-analyses that did achieve statistical significance, and in our view it would be premature to conclude that this effect does not exist.

These examples illustrate the difficulty in deciding whether conducting a power analysis is appropriate. Even tiny effect sizes could hypothetically still exist: in any biological system the probability that an effect is precisely null is itself zero – therefore all effects "exist" by this definition (with certain exceptions, e.g. in the context of randomization), even if to detect them we might need to test more individuals than are currently alive. However, the notion of "falsely rejecting the null hypothesis" then loses its meaning (Jacob Cohen, 1994). One approach would be to assume that an effect does not exist until the observed evidence suggests that the null hypothesis can be rejected, consistent with the logical basis of classical statistical inference. This would avoid any potential bias towards very low power estimates due to non-existent effects. On the other hand, this approach raises the potential problem of excluding effects that are genuinely very small, which may cause a bias in the other direction. Within the constraints of the null hypothesis significance testing framework, it is impossible to be confident that an effect does not exist at all. Therefore, we cannot simply assume an effect does not exist after failing to reject the null hypothesis, since a small effect could go undetected.

Motivated by this logic, we initially included studies from ‘null meta-analyses’ (i.e. where the estimated effect size from the meta-analysis was not significantly different from the null at the conventional alpha=0.05) in our GMMs (Figure 3a). However, we note that excluding power statistics from studies included in null meta-analyses may provide an overestimation of power, because in many instances there remains uncertainty as to whether or not a true effect exists. Nonetheless, with the above caveats in mind, we also wished to assess the degree to which null meta-analyses may have impacted the results. Null results occurred in 7 of the 49 meta-analyses (92 of the 730 individual studies), contributing a substantial proportion of the extremely low powered studies (<10% power; Figure 3a, white pie chart segment of C1). When we restricted our analysis only to studies within meta-analyses that reported statistically significant results (‘non-null’ meta-analyses), the median study power (unsurprisingly) increased, but only slightly, to 30%, and the nature of the resulting GMM distribution did not change substantially (see Figure 3b). Thus, excluding null meta-analyses does not provide a radically different picture. Therefore, we also examined another potential contributor to power variability in neuroscience: the influence of specific subfields of neuroscience.

Power in neuroscience subfields

As described above, we categorised each meta-analysis into one of six methodological subfields. Interestingly, statistical power varied significantly according to subfield (permutation test of equivalence: $p<0.001$), with genetic association studies lower (11% median power) than any other subfield examined (all Mann-Whitney U tests $p<0.001$). This is consistent with the original report by Button and colleagues, which reported the median power of animal studies (18% and 31% for two meta-analyses) and structural brain imaging studies (8% across 41 meta-analyses). However, even within specific subfields, the distribution of power is multimodal (see Figure 3c-h). This could represent variability in statistical practices across studies, but another possible explanation is that the size of the effect being studied varies substantially between meta-analyses, even within the same subfield. This alternative explanation may, at least in part, account for the variability between (and within) subfields of neuroscience.

The large number of extremely low powered candidate gene association studies warrants additional comment. These were included in the original analysis because the Web of Science classifies such studies as “neuroscience” if the phenotypes in question are neurological or psychiatric disorders. However, modern genome-wide association studies have revealed that the overwhelming majority of candidate gene association studies have been underpowered, because the reliable associations that have been identified are extremely small (Flint and Munafò, 2013); thus, very low power is expected within this subgroup, which our analysis confirms (see Figure 3c). This subgroup of studies can offer important lessons to the rest of neuroscience: without large genetic consortia, the field of neuropsychiatric genetics might still be labouring under the misapprehension that individual common variants make substantial contributions to the risk for developing disorders. Providing that sampling and measurement are standardised, pooling data across multiple sites has the potential to improve dramatically not only statistical power, but also the precision on estimates of effect size.

Since numerous studies report that candidate gene association studies are severely underpowered (Klerk et al., 2002; Colhoun et al., 2003; Duncan and Keller, 2011), and given that candidate gene association studies comprised over one-third of our total sample of studies, we suspected that they might contribute heavily to the lowest-power peak in our distribution. We confirmed this: in the absence of genetic studies, many studies remained underpowered, but the distribution contained proportionally fewer studies in the lowest-power peak (around 10% power) (Figure 4a). Although low power is clearly not limited to candidate gene association studies, they nonetheless seem to have a greater influence on the overall power distribution than any other subfield, skewing the distribution towards the lowest-power peak (Figure 4b-f).

Figure 4. Gaussian Mixture Models (GMMs) excluding each subfield.

GMMs for the whole population of studies excluding genetic studies (a), excluding psychology studies (b), excluding neurochemistry studies (c), excluding treatment studies (d), excluding imaging studies (e), and excluding the remaining miscellaneous studies (f). Compare with the distribution including all studies (Figure 3a).

Estimations of effect size

An important factor contributing to the estimation of power is whether the effect size was estimated accurately *a priori*. If researchers initially overestimated the effect size, even the sample size specified by a power calculation would be insufficient to detect a real, but smaller effect. Interestingly, our analysis also shows the existence of very high powered studies within neuroscience, in which far more subjects have been included than would technically be warranted by a power analysis. In this case, an *a priori* underestimate of effect size could yield a very high powered study, if an effect proves to be larger than initially expected (which has occasionally been reported (Open Science Collaboration, 2015)). Another important consideration is that an over-estimation of effect size might occur due to publication bias, which will skew effect size estimates from meta-analyses upwards, resulting in an optimistic power estimate. This is an important caveat to the results we report here: a bias toward publishing significant results means that the power estimates we report will represent upper bounds on the true power statistics. Unfortunately, we could not adequately address this potential confound directly, since tests of publication bias themselves have very low power, particularly if the number of studies in a meta-analysis is low. However, publication bias has long been reported in psychology (Francis, 2012) and neuroscience (Sena et al., 2010), so it is reasonable to assume that it has inflated estimates of statistical power in these analyses.

Simulating power in hypothetical fields

One clear conclusion of our analyses is that the interplay between the proportion of true effects and the power to detect those effects is crucial in determining the power distribution of a field. We simulated four power graphs for hypothetical fields to illustrate this point: one with low power (~50%), but where all effects exist (Figure 5a); one with high power (~90%), where all effects exist (Figure 5b); one with low power (~50%), where only a minority (25%) of effects exist (Figure 5c); and high power (~90%), but where only a minority (25%) of effects exist (Figure 5d). We found that the 'low power' field did not resemble the distribution of power in neuroscience we observed (Figure 3a). Instead, our findings were closest to a mixture of two distributions: Figure 5c, with low (~50%) power, and where only 25% of findings are true effects; and Figure 5d, with high (~90%) power, but where only 25% of findings are true effects. This would be consistent with the notion that the absence of true effects may contribute to the distribution of statistical power in neuroscience.

Figure 5. Simulated power distributions for four hypothetical fields. (a) 'Easy field' with low power (~0.5) and all effects exist; (b) 'Easy field' with high power (~0.9) and all effects exist; (c) 'Hard

field' with low power (~0.5) (for those effects that exist), but where effects exist in only 25% of cases; (d) 'Hard field' with high power (~0.9) (for those effects that exist), but where effects exist in only exist in 25% of cases. Power distributions were simulated by generating 50,000 samples with fixed sample-size (N=45) while varying effect-size. For each panel, the effect-size was sampled from a truncated (effect-size>0) Gaussian distribution with mean 0.3 (a & c) or 0.49 (b & d), so as to represent low or high power respectively. For the 'hard' fields (c & d), 75% of the effect-size sample was generated from a half-Gaussian distribution with mean=0. SD was set to 0.07 for all effect size distributions. Similar results can be obtained by fixing the effect size and varying the sample size.

Discussion

Implications for neuroscience

We argue that a very influential analysis (cited over 1500 times at the time of writing) does not adequately describe the full variety of statistical power in neuroscience. Our analyses show that the dataset is insufficiently characterized by a single distribution. Instead, power varies considerably, including between subfields of neuroscience, and is particularly low for candidate gene association studies. Conducting power analyses for null effects may also contribute to low estimates in some cases, though determining when this has occurred is challenging. Importantly, however, power is far from adequate in every subfield.

Our analyses do not negate the importance of the original work in highlighting poor statistical practice in the field, but they do reveal a more nuanced picture. In such a diverse field as neuroscience, it is not surprising that statistical practices differ. While Button and colleagues were careful to point out that they identified a range of powers in neuroscience, their reporting of a median result could be interpreted as implying that the results were drawn from a single distribution, which our analyses suggest is not the case. We confirm that low power is clearly present in many studies, and agree that focusing on power is a critical step in improving the replicability and reliability of findings in neuroscience. However, we also argue that low statistical power in neuroscience is neither consistent nor universal.

Ethical issues accompany both under- and over-powered studies. Animal sacrifices, drugs taken to human trials, and government funding are all wasted if power is too low. However, blindly increasing sample size across the board, simply to satisfy concerns about field-wide power failures, is also not the best use of resources. Instead, each study design needs to be considered on its own merits. In this vein, one response to the original article pointed out that any measure of a study's projected value suffers from diminishing marginal returns: every additional animal or human participant adds less statistical value than the previous one (Bacchetti et al., 2005, 2008; 2013).

Studies with extremely large sample sizes can also fall prey to statistically significant findings for trivial effects that are unlikely to be either theoretically or clinical important (Lenth, 2001; Ioannidis, 2005; Friston, 2012; Quinlan, 2013). In other words, the assessment of power is determined by what we consider to be an interesting (i.e. nontrivial) effect size (Cohen, 1988). This dependency means that power considerations are meaningless in the absence of assumptions about how large effect sizes need to be in order to be considered theoretically or clinically important; and this may vary dramatically across different fields. This is particularly relevant in fields where multiple comparisons are performed routinely, such as genetics and neuroimaging (Friston, 2012). Conversely, smaller studies can only detect large effect sizes, and may suffer from imprecise estimates of effect size and interpretive difficulties. Crucially, there is no single study design that will optimise power for every genetic locus or brain area. In fact, power estimates for individual studies are themselves extremely noisy and may say little about the actual power in any given study. However, a move away from presenting only p-values and towards reporting point estimates and confidence intervals (as long advocated by statisticians), and towards sharing data to improve such estimates, would allow researchers to make better informed decisions about whether an effect is likely to be clinically or theoretically useful.

Conclusion

We have demonstrated the great diversity of statistical power in neuroscience. Do our findings lessen concerns about statistical power in neuroscience? Unfortunately not. In fact, the finding that the distribution of power is highly heterogeneous demonstrates an undesirable inconsistency, both within and between methodological subfields. Yet within this variability are several appropriately, and even very high powered studies. Therefore, we should not tar all studies with the same brush, but instead encourage investigators to engage in the best research practices, including preregistration of study protocols (ensuring the study will have sufficient power), routine publication of null results, and avoiding practices such as p-hacking that lead to biases in the published literature.

References

Babbage DR, Yim J, Zupan B, Neumann D, Tomita MR, Willer B (2011) Meta-analysis of facial affect recognition difficulties after traumatic brain injury. Neuropsychology 25:277.

Bacchetti P (2013) Small sample size is not the real problem. Nat Rev Neurosci 14:585–585.

Bacchetti P, McCulloch CE, Segal MR (2008) Simple, defensible sample sizes based on cost efficiency. Biometrics 64:577–585.

Bacchetti P, Wolf LE, Segal MR, McCulloch CE (2005) Ethics and sample size. Am J Epidemiol 161:105–110.

Bai H (2011) Meta-analysis of 5, 10-methylenetetrahydrofolate reductase gene poymorphism as a risk factor for ischemic cerebrovascular disease in a Chinese Han population. Neural Regen Res 6:277–285.

Barch DM, Yarkoni T (2013) Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. Cogn Affect Behav Neurosci 13:687–689.

Bishop CM (2006) Pattern recognition and machine learning. springer New York.

Björkhem-Bergman L, Asplund AB, Lindh JD (2011) Metformin for weight reduction in non-diabetic patients on antipsychotic drugs: a systematic review and meta-analysis. J Psychopharmacol (Oxf) 25:299–305.

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. Bayesian Anal 1:121–143.

Bucossi S, Ventriglia M, Panetta V, Salustri C, Pasqualetti P, Mariani S, Siotto M, Rossini PM, Squitti R (2011) Copper in Alzheimer’s disease: a meta-analysis of serum, plasma, and cerebrospinal fluid studies. J Alzheimers Dis 24:175–185.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Chamberlain SR, Robbins TW, Winder-Rhodes S, Müller U, Sahakian BJ, Blackwell AD, Barnett JH (2011) Translational approaches to frontostriatal dysfunction in attention-deficit/hyperactivity disorder using a computerized neuropsychological battery. Biol Psychiatry 69:1192–1203.

Chang W, Arfken CL, Sangal MP, Boutros NN (2011a) Probing the relative contribution of the first and second responses to sensory gating indices: A meta-analysis. Psychophysiology 48:980–992.

Chang X-L, Mao X-Y, Li H-H, Zhang J-H, Li N-N, Burgunder J-M, Peng R, Tan E-K (2011b) Functional parkin promoter polymorphism in Parkinson’s disease: new data and meta-analysis. J Neurol Sci 302:68–71.

Chen C, Xu T, Chen J, Zhou J, Yan Y, Lu Y, Wu S (2011) Allergy and risk of glioma: a meta-analysis. Eur J Neurol 18:387–395.

Chung AK, Chua S (2011) Effects on prolongation of Bazett’s corrected QT interval of seven second-generation antipsychotics in the treatment of schizophrenia: a meta-analysis. J Psychopharmacol (Oxf) 25:646–666.

Cohen J (1988) Statistical power analysis for the behavioral sciences. Vol. 2. Lawrence Earlbaum Assoc Hillsdale NJ.

Corduneanu A, Bishop CM (2001) Variational Bayesian model selection for mixture distributions. In, pp 27–34. Morgan Kaufmann Waltham, MA.

Domellöf E, Johansson A-M, Rönnqvist L (2011) Handedness in preterm born children: a systematic review and a meta-analysis. Neuropsychologia 49:2299–2310.

Etminan N, Vergouwen MD, Ilodigwe D, Macdonald RL (2011) Effect of pharmaceutical treatment on vasospasm, delayed cerebral ischemia, and clinical outcome in patients with aneurysmal subarachnoid hemorrhage: a systematic review and meta-analysis. J Cereb Blood Flow Metab 31:1443–1451.

Feng X, Wang F, Zou Y, Li W, Tian Y, Pan F, Huang F (2011) Association of FK506 binding protein 5 (FKBP5) gene rs4713916 polymorphism with mood disorders: a meta-analysis. Acta Neuropsychiatr 23:12–19.

Flint J, Munafò MR (2013) Candidate and non-candidate genes in behavior genetics. Curr Opin Neurobiol 23:57–61.

Francis G (2012) Too good to be true: Publication bias in two prominent studies from experimental psychology. Psychon Bull Rev 19:151–156.

Friston K (2012) Ten ironic rules for non-statistical reviewers. Neuroimage 61:1300–1310.

Fryar C, Gu Q, Ogden (2012) Anthropometric Reference Data for Children and Adults: United States, 2007-2010. U.S. Department of Health and Human Services.

Gershman SJ, Blei DM (2012) A tutorial on Bayesian nonparametric models. J Math Psychol 56:1–12.

Green M, Matheson S, Shepherd A, Weickert C, Carr V (2011) Brain-derived neurotrophic factor levels in schizophrenia: a systematic review with meta-analysis. Mol Psychiatry 16:960–972.

Han X-M, Wang C-H, Sima X, Liu S-Y (2011) Interleukin-6− 174G/C polymorphism and the risk of Alzheimer’s disease in Caucasians: A meta-analysis. Neurosci Lett 504:4–8.

Hannestad J, DellaGioia N, Bloch M (2011) The effect of antidepressant medication treatment on serum levels of inflammatory cytokines: a meta-analysis. Neuropsychopharmacology 36:2452–2459.

Hua Y, Zhao H, Kong Y, Ye M (2011) Association between the MTHFR gene and Alzheimer's disease: a meta-analysis. Int J Neurosci 121:462–471.

Hyndman RJ (1996) Computing and graphing highest density regions. Am Stat 50:120–126.

Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124.

Jacob Cohen (1994) The earth is round (p<0.05). Am Psychol 49:997–1003.

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795.

Kurihara K, Welling M, Teh YW (2007) Collapsed Variational Dirichlet Process Mixture Models. In, pp 2796–2801.

Lenth RV (2001) Some practical guidelines for effective sample size determination. Am Stat 55:187–193.

Lindson N, Aveyard P (2011) An updated meta-analysis of nicotine preloading for smoking cessation: investigating mediators of the effect. Psychopharmacology (Berl) 214:579–592.

Liu H, Liu M, Wang Y, Wang X-M, Qiu Y, Long J-F, Zhang S-P (2011a) Association of 5-HTT gene polymorphisms with migraine: a systematic review and meta-analysis. J Neurol Sci 305:57–66.

Liu J, Sun Q, Tang B, Hu L, Yu R, Wang L, Shi C, Yan X, Pan Q, Xia K (2011b) PITX3 gene polymorphism is associated with Parkinson's disease in Chinese population. Brain Res 1392:116–120.

Lubke GH, Muthén B (2005) Investigating population heterogeneity with factor mixture models. Psychol Methods 10:21.

MacKillop J, Amlung MT, Few LR, Ray LA, Sweet LH, Munafò MR (2011) Delayed reward discounting and addictive behavior: a meta-analysis. Psychopharmacology (Berl) 216:305–321.

Maneeton N, Maneeton B, Srisurapanont M, Martin SD (2011) Bupropion for adults with attention-deficit hyperactivity disorder: Meta-analysis of randomized, placebo-controlled trials. Psychiatry Clin Neurosci 65:611–617.

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press.

Nosek BA, Spies JR, Motyl M (2012) Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci 7:615–631.

Ohi K, Hashimoto R, Yasuda Y, Fukumoto M, Yamamori H, Umeda-Yano S, Kamino K, Ikezawa K, Azechi M, Iwase M (2011) The SIGMAR1 gene is associated with a risk of schizophrenia and activation of the prefrontal cortex. Prog Neuropsychopharmacol Biol Psychiatry 35:1309–1315.

O'Keefe DJ (2007) Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. Commun Methods Meas 1:291–299.

Olabi B, Ellison-Wright I, McIntosh AM, Wood SJ, Bullmore E, Lawrie SM (2011) Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. Biol Psychiatry 70:88–96.

Oldershaw A, Hambrook D, Stahl D, Tchanturia K, Treasure J, Schmidt U (2011) The socio-emotional processing stream in anorexia nervosa. Neurosci Biobehav Rev 35:970–988.

Oliver BJ, Kohli E, Kasper LH (2011) Interferon therapy in relapsing-remitting multiple sclerosis: a systematic review and meta-analysis of the comparative trials. J Neurol Sci 302:96–105.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349:aac4716.

Peerbooms OL, van Os J, Drukker M, Kenis G, Hoogveld L, De Hert M, Delespaul P, van Winkel R, Rutten BP (2011) Meta-analysis of MTHFR gene variants in schizophrenia, bipolar disorder and unipolar depressive disorder: evidence for a common genetic vulnerability? Brain Behav Immun 25:1530–1543.

Pizzagalli DA (2011) Frontocingulate dysfunction in depression: toward biomarkers of treatment response. Neuropsychopharmacology 36:183–206.

Quinlan PT (2013) Misuse of power: in defence of small-scale science. Nat Rev Neurosci 14:585–585.

Rist PM, Diener H-C, Kurth T, Schürks M (2011) Migraine, migraine aura, and cervical artery dissection: a systematic review and meta-analysis. Cephalalgia 31:886–896.

Samworth R, Wand M (2010) Asymptotics and optimal bandwidth selection for highest density region estimation. Ann Stat 38:1767–1792.

Sena ES, Van Der Worp HB, Bath PM, Howells DW, Macleod MR (2010) Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol 8:e1000344.

Sexton CE, Kalu UG, Filippini N, Mackay CE, Ebmeier KP (2011) A meta-analysis of diffusion tensor imaging in mild cognitive impairment and Alzheimer’s disease. Neurobiol Aging 32:2322-e5.

Shum D, Levin H, Chan RC (2011) Prospective memory in patients with closed head injury: a review. Neuropsychologia 49:2156–2165.

Sim H, Shin B-C, Lee MS, Jung A, Lee H, Ernst E (2011) Acupuncture for carpal tunnel syndrome: a systematic review of randomized controlled trials. J Pain 12:307–314.

Song F, Poljak A, Valenzuela M, Mayeux R, Smythe GA, Sachdev PS (2011) Meta-analysis of plasma amyloid-β levels in Alzheimer’s disease. J Alzheimers Dis 26:365–375.

Sun Q, Fu Y, Sun A, Shou Y, Zheng M, Li X, Fan D (2011) Correlation of E-selectin gene polymorphisms with risk of ischemic stroke A meta-analysis. Neural Regen Res 6.

Tian Y, Kang L, Wang H, Liu Z (2011) Meta-analysis of transcranial magnetic stimulation to treat post-stroke dysfunction. Neural Regen Res 6.

Trzesniak C, Kempton MJ, Busatto GF, de Oliveira IR, Galvao-de Almeida A, Kambeitz J, Ferrari MCF, Santos Filho A, Chagas MH, Zuardi AW (2011) Adhesio interthalamica alterations in schizophrenia spectrum disorders: A systematic review and meta-analysis. Prog Neuropsychopharmacol Biol Psychiatry 35:877–886.

Veehof MM, Oskam M-J, Schreurs KM, Bohlmeijer ET (2011) Acceptance-based interventions for the treatment of chronic pain: a systematic review and meta-analysis. PAIN® 152:533–542.

Vergouwen MD, Etminan N, Ilodigwe D, Macdonald RL (2011) Lower incidence of cerebral infarction correlates with improved functional outcome after aneurysmal subarachnoid hemorrhage. J Cereb Blood Flow Metab 31:1545–1553.

Vieta E, Günther O, Locklear J, Ekman M, Miltenburger C, Chatterton ML, Åström M, Paulsson B (2011) Effectiveness of psychotropic medications in the maintenance phase of bipolar disorder: a meta-analysis of randomized controlled trials. Int J Neuropsychopharmacol 14:1029–1049.

Wand MP, Marron JS, Ruppert D (1991) Transformations in density estimation. J Am Stat Assoc 86:343–353.

Wisdom NM, Callahan JL, Hawkins KA (2011) The effects of apolipoprotein E on non-impaired cognitive functioning: a meta-analysis. Neurobiol Aging 32:63–74.

Witteman J, van IJzendoorn MH, van de Velde D, van Heuven VJ, Schiller NO (2011) The nature of hemispheric specialization for linguistic and emotional prosodic perception: a meta-analysis of the lesion literature. Neuropsychologia 49:3722–3738.

Woon F, Hedges DW (2011) Gender does not moderate hippocampal volume deficits in adults with posttraumatic stress disorder: A meta-analysis. Hippocampus 21:243–252.

Xuan C, Zhang B-B, Li M, Deng K-F, Yang T, Zhang X-E (2011) No association between APOE epsilon 4 allele and multiple sclerosis susceptibility: a meta-analysis from 5472 cases and 4727 controls. J Neurol Sci 308:110–116.

Yang W, Kong F, Liu M, Hao Z (2011a) Systematic review of risk factors for progressive ischemic stroke. Neural Regen Res 6:346–352.

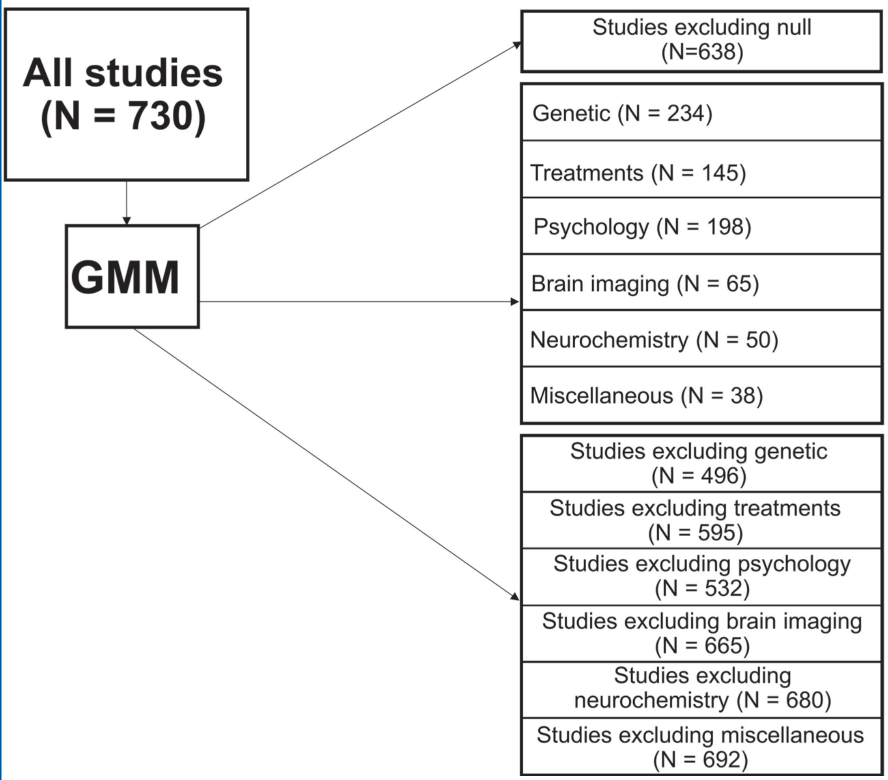
Yang Z, Li W, Huang T, Chen J, Zhang X (2011b) Meta-analysis of Ginkgo biloba extract for the treatment of Alzheimer’s disease. Neural Regen Res 6:1125–1129.

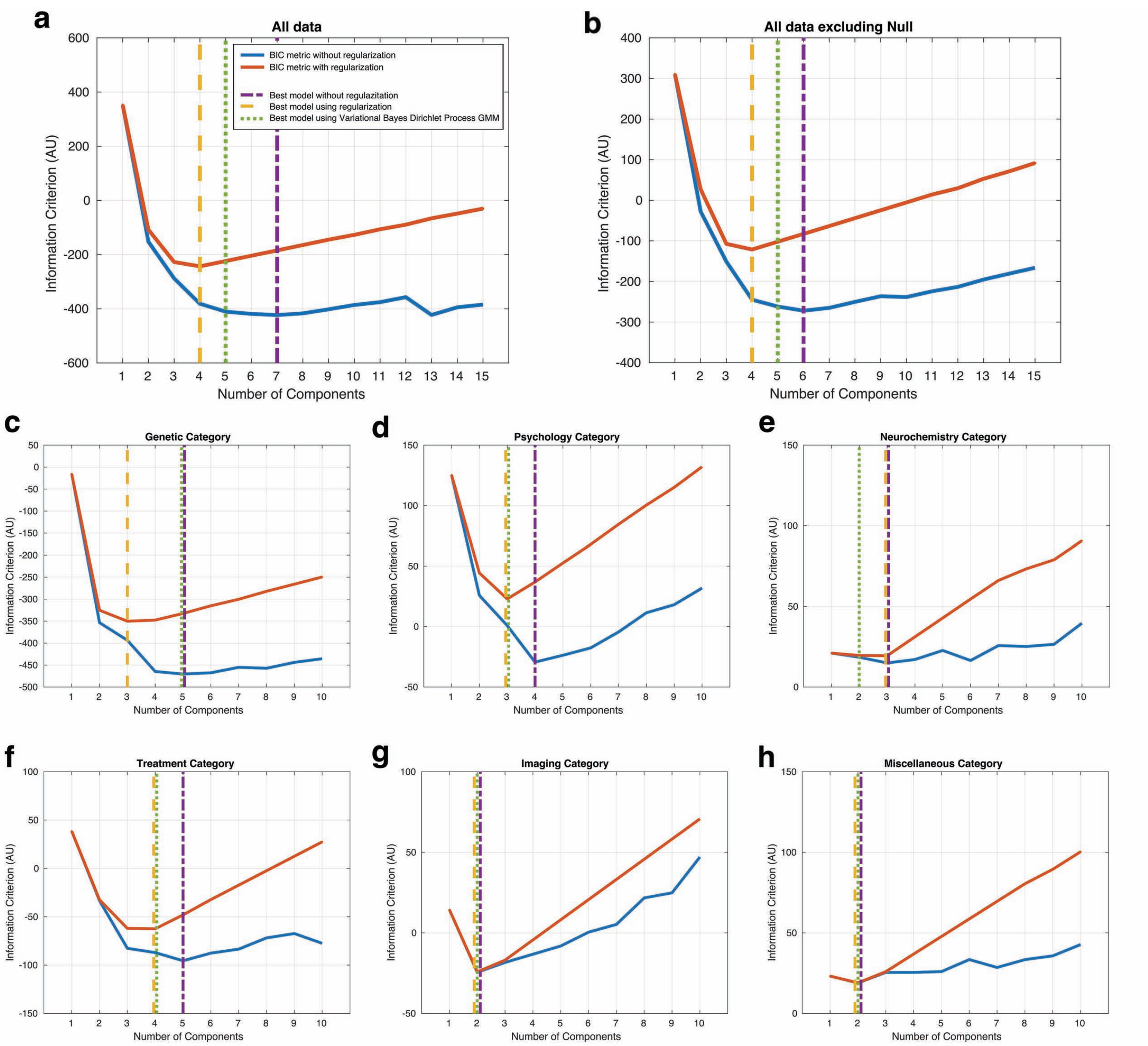
Yuan H, Yang X, Kang H, Cheng Y, Ren H, Wang X (2011) Meta-analysis of tau genetic polymorphism and sporadic progressive supranuclear palsy susceptibility. Neural Regen Res 6:353–359.

Zafar SN, Iqbal A, Farez MF, Kamatkar S, de Moya MA (2011) Intensive insulin therapy in brain injury: a meta-analysis. J Neurotrauma 28:1307–1317.

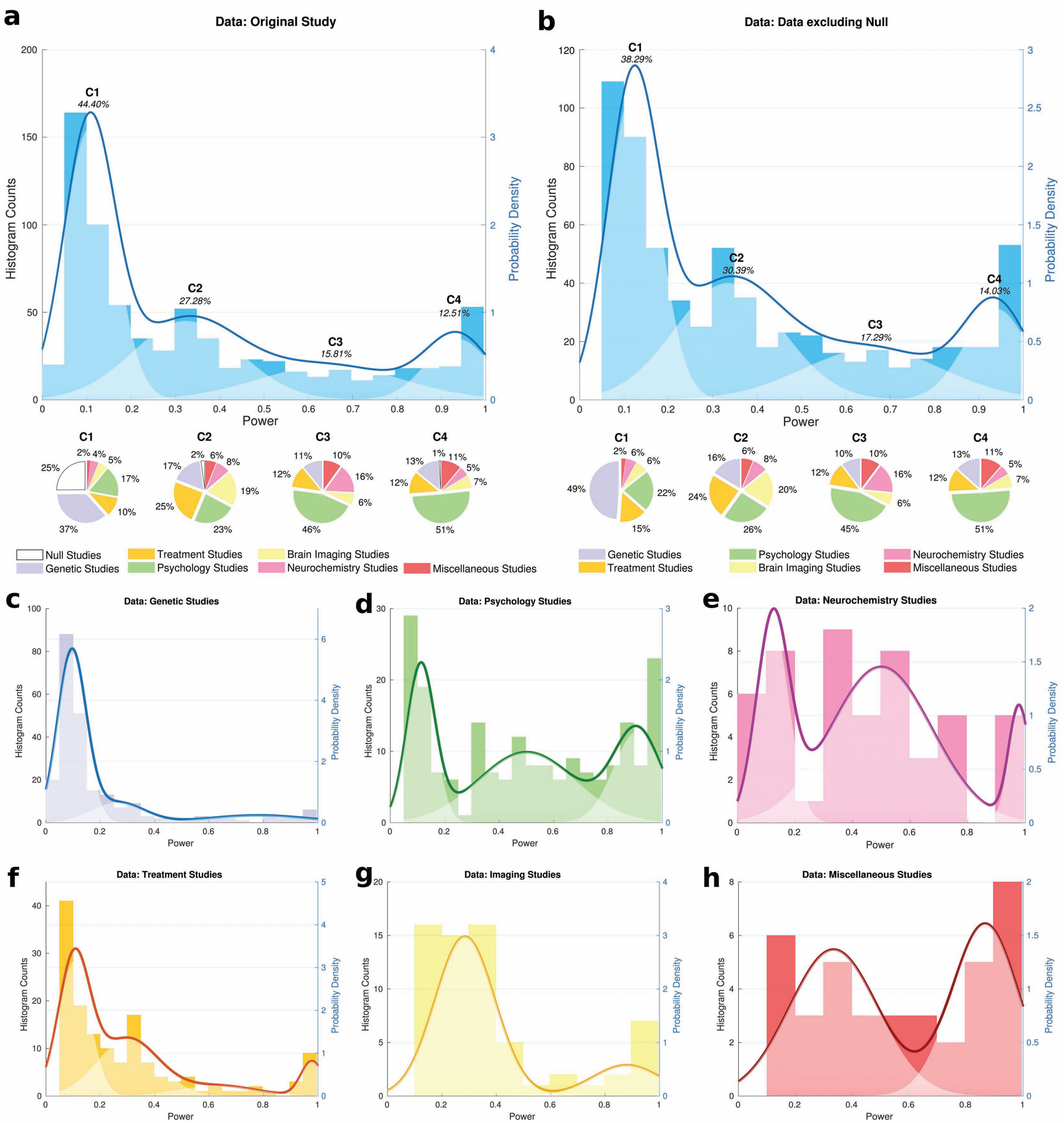
Zhang Y, Zhang J, Tian C, Xiao Y, Li X, He C, Huang J, Fan H (2011) The− 1082G/A polymorphism in IL-10 gene is associated with risk of Alzheimer’s disease: a meta-analysis. J Neurol Sci 303:133–138.

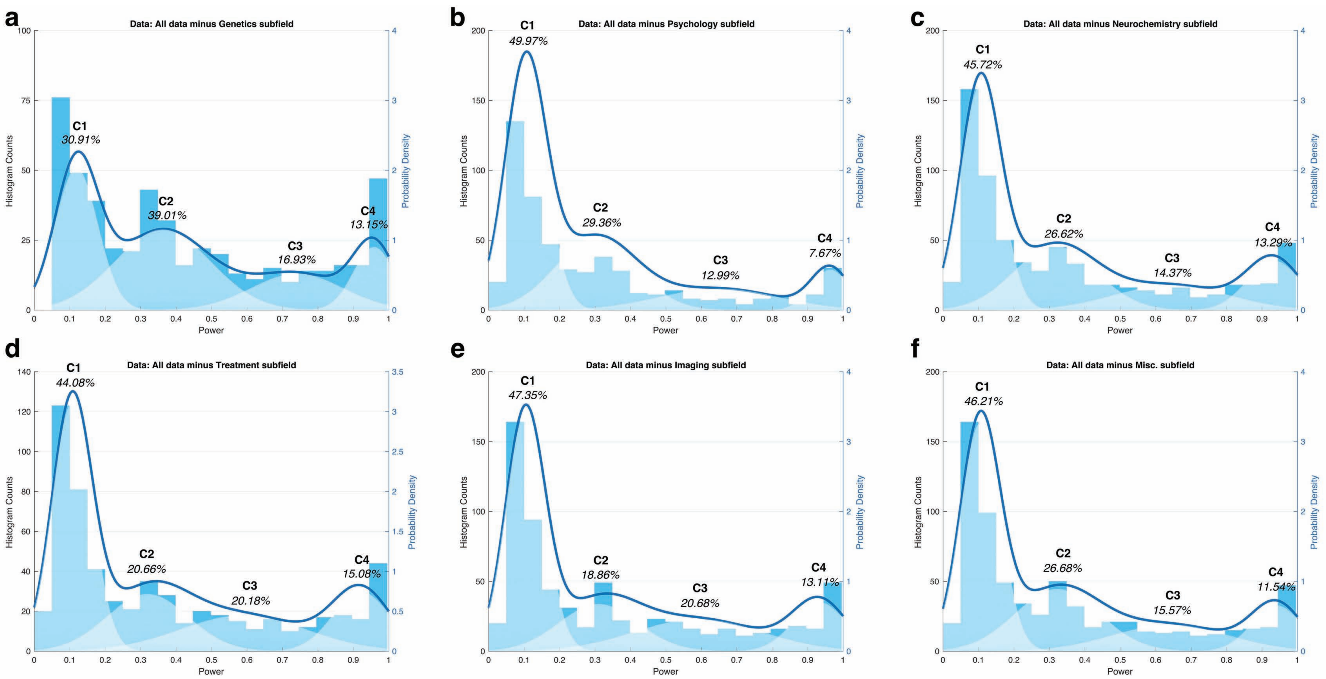
Zhu Y, He Z-Y, Liu H-N (2011) Meta-analysis of the relationship between homocysteine, vitamin B 12, folate, and multiple sclerosis. J Clin Neurosci 18:933–938.

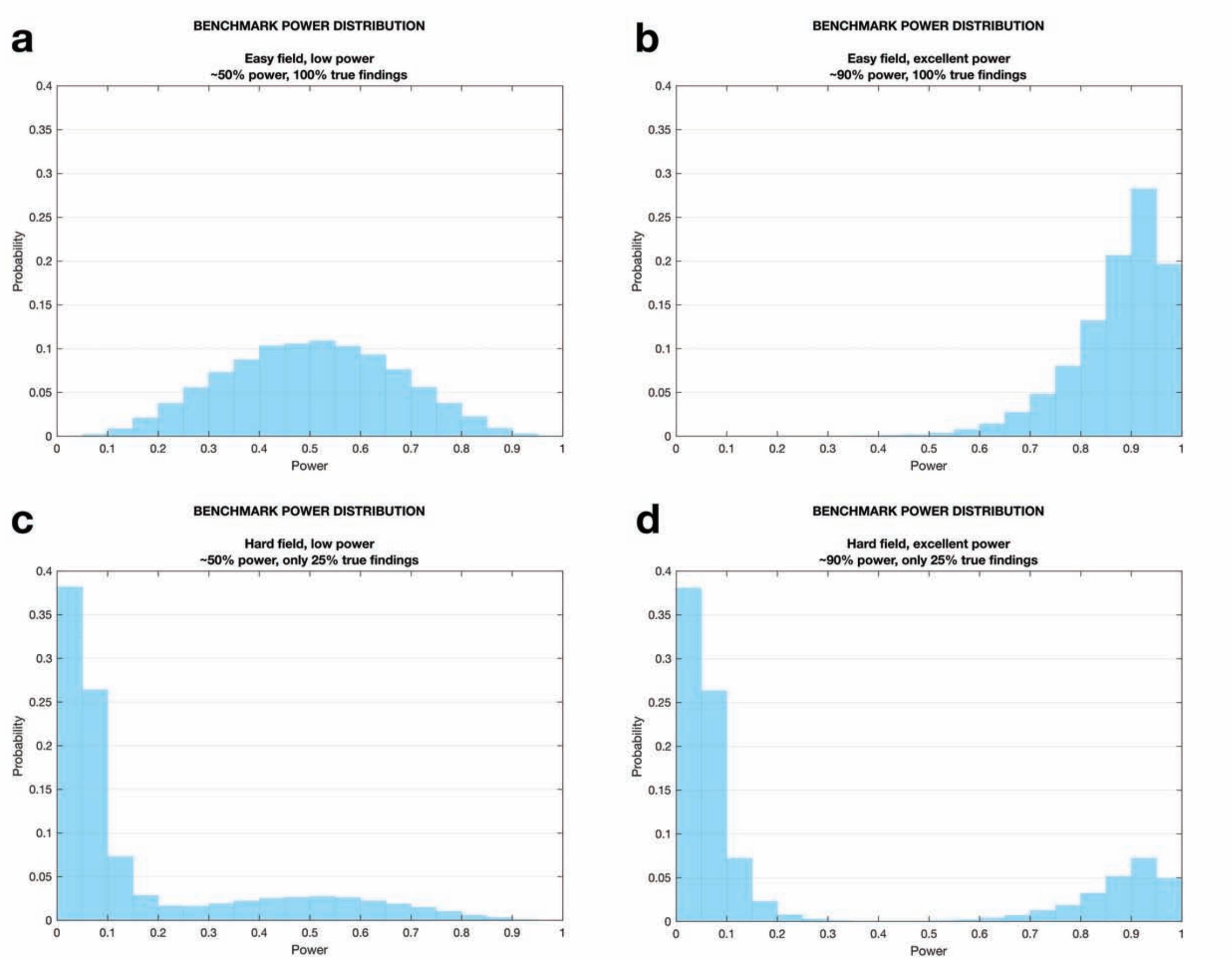

All studies
(N = 730)
GMM
Studies excluding null
(N=638)
Genetic (N = 234)
Treatments (N = 145)
Psychology (N = 198)
Brain imaging (N = 65)
Neurochemistry (N = 50)
Miscellaneous (N = 38)
Studies excluding genetic
(N = 496)
Studies excluding treatments
(N = 595)
Studies excluding psychology
(N = 532)
Studies excluding brain imaging
(N = 665)
Studies excluding
neurochemistry (N = 680)
Studies excluding miscellaneous
(N = 692)



a
All data
BIC metric without regularization
BIC metric with regularization
Best model without regulazitation
Best model using regularization
Best model using Variational Bayes Dirichlet Process GMM
Information Criterion (AU)
Number of Components
b
All data excluding Null
Information Criterion (AU)
Number of Components
c
Genetic Category
d
Psychology Category
e
Neurochemistry Category
f
Treatment Category
g
Imaging Category
h
Miscellaneous Category


a
Data: Original Study
b
Data: Data excluding Null
c
Data: Genetic Studies
d
Data: Psychology Studies
e
Data: Neurochemistry Studies
f
Data: Treatment Studies
g
Data: Imaging Studies
h
Data: Miscellaneous Studies
Histogram Counts
Probability Density
Power
C1 44.40%
C2 27.28%
C3 15.81%
C4 12.51%
C1 38.29%
C2 30.39%
C3 17.29%
C4 14.03%
Null Studies
Genetic Studies
Treatment Studies
Psychology Studies
Brain Imaging Studies
Neurochemistry Studies
Miscellaneous Studies


a
Data: All data minus Genetics subfield
C1
30.91%
C2
39.01%
C3
16.93%
C4
13.15%
Histogram Counts
Probability Density
Power
b
Data: All data minus Psychology subfield
C1
49.97%
C2
29.36%
C3
12.99%
C4
7.67%
Histogram Counts
Probability Density
Power
c
Data: All data minus Neurochemistry subfield
C1
45.72%
C2
26.62%
C3
14.37%
C4
13.29%
Histogram Counts
Probability Density
Power
d
Data: All data minus Treatment subfield
C1
44.08%
C2
20.66%
C3
20.18%
C4
15.08%
Histogram Counts
Probability Density
Power
e
Data: All data minus Imaging subfield
C1
47.35%
C2
18.86%
C3
20.68%
C4
13.11%
Histogram Counts
Probability Density
Power
f
Data: All data minus Misc. subfield
C1
46.21%
C2
26.68%
C3
15.57%
C4
11.54%
Histogram Counts
Probability Density
Power


a
BENCHMARK POWER DISTRIBUTION
Easy field, low power
~50% power, 100% true findings
b
BENCHMARK POWER DISTRIBUTION
Easy field, excellent power
~90% power, 100% true findings
c
BENCHMARK POWER DISTRIBUTION
Hard field, low power
~50% power, only 25% true findings
d
BENCHMARK POWER DISTRIBUTION
Hard field, excellent power
~90% power, only 25% true findings
Probability
Power