Estimating the prevalence of missing experiments in a neuroimaging meta-analysis

Pantelis Samartsidis[1], Silvia Montagna[2], Angela R. Laird[3], Peter T. Fox[4,5], Timothy D. Johnson[6] and Thomas E. Nichols[7]

[1]*MRC Biostatistics Unit, University of Cambridge* [2]*Dipartimento di Scienze Economico-sociali e Matematico-statistiche (ESOMAS), University of Torino* [3]*Department of Physics, Florida International University* [4]*Research Imaging Institute, University of Texas at San Antonio* [5]*South Texas Veterans Health Care System* [6]*Department of Biostatistics, University of Michigan* [7]*Oxford Big Data Institute, University of Oxford*

August 24, 2020

Abstract

Coordinate-based meta-analyses (CBMA) allow researchers to combine the results from multiple fMRI experiments with the goal of obtaining results that are more likely to generalise. However, the interpretation of CBMA findings can be impaired by the file drawer problem, a type of publications bias that refers to experiments that are carried out but are not published. Using foci per contrast count data from the BrainMap database, we propose a zero-truncated modelling approach that allows us to estimate the prevalence of non-significant experiments. We validate our method with simulations and real coordinate data generated from the Human Connectome Project. Application of our method to the data from BrainMap provides evidence for the existence of a file drawer effect, with the rate of missing experiments estimated as at least 6 per 100 reported. The R code that we used is available at https://osf.io/ayhfv/.

1 Introduction

Now over 25 years old, functional magnetic resonance imaging (fMRI) has made significant contributions in improving our understanding of the human brain function. However, the inherent limitations of fMRI experiments have raised concerns regarding the validity and replicability of findings [1]. These limitations include poor test-retest reliability [2], excess of false positive findings [3] and small sample sizes [4]. Meta-analyses play an important role in the field of task-activation fMRI as they provide a means to address the aforementioned problems by synthesising the results from multiple experiments and thus draw more reliable conclusions. Since the overwhelming majority of authors rarely share the full data, *coordinate-based meta-analyses* (CBMA), which use the $x - y - z$ coordinates (foci) of peak activations that are typically published, are the main approach for the meta-analysis of task-activation fMRI data.

As in any meta-analysis, the first step in a CBMA is a literature search. During this step investigators use databases to retrieve all previous work which is relevant to the question of interest [5]. Ideally, this process will yield an exhaustive or at least representative sample of studies on a specific topic. Unfortunately, literature search is subject to the *file drawer* problem [6, 7]. This problem refers to research studies that are initiated but are not published. When these studies are missing at random (i.e. the reasons that they remain unpublished are independent of their findings), then the pool of studies reduces but the results of a meta-analysis remain reliable. However, if the censoring relates to the findings of a study (e.g. due to rejection by journals reluctant to publish negative results), then meta-analyses may yield biased estimates of the effect of interest [8, 9].

In CBMA, the unit of observation is a *contrast/experiment* (these terms are used interchangeably throughout the paper) and not a *study/paper* (these terms are also used interchangeably throughout the paper), because the latter may include multiple contrasts that can be used in a single meta-analysis.

Hence the file drawer includes contrasts that find no significant activation clusters i.e. ones that report no foci. Such experiments often remain unpublished because when writing a paper, authors focus on the other, significant experiments that they conducted. Moreover, even if mentioned in the final publication, these contrasts are typically not mentioned in the table of foci, and are not registered in the databases which researchers use to retrieve data for their CBMA. The bias introduced by not considering contrasts with no foci depends on how often these occur in practice. For example, if only 1 out of 100 contrasts is null then not considering zero-count contrasts is unlikely to have an impact on the results of a CBMA. However, if this relative proportion is high, then the findings of a CBMA will be misleading in that they will overestimate the effect of interest.

Some authors have attempted to assess the evidence for the existence of publication biases in the field of fMRI. One example is [10], who found evidence for publication biases in 74 studies of tasks involving working memory. The authors used the maximum test statistic reported in the frontal lobe as the effect estimate in their statistical tests. Another example is [11], who studied the relation between sample size and the total number of activations and reached similar conclusions as [10]. However, to date there has been no work on estimating a fundamental file drawer quantity, that is the prevalence of null experiments.

In this paper, we propose a model for estimating the prevalence of zero-count contrasts in the context of CBMA. Our approach is outlined in Figure 1. Let the *sampling frame* be all K neuroimaging experiments of interest that were completed, published or not, where each element of the sampling frame is a statistic map for a contrast. For any contrast in the sampling frame, let $\pi(n|\boldsymbol{\theta})$ be the probability mass function of the number of foci per contrast, where $\boldsymbol{\theta}$ is a vector of parameters. Hence, null contrasts occur with probability $p_0 = \pi(0 \mid \boldsymbol{\theta})$ and there are Kp_0 in total. These are unavailable for meta-analysis due to lack of significance. However, the remaining $k = K(1 - p_0)$ significant experiments are published and are available to draw inference regarding $\boldsymbol{\theta}$. This allows us estimate p_0 and thus the prevalence of zero-count contrasts in the population.

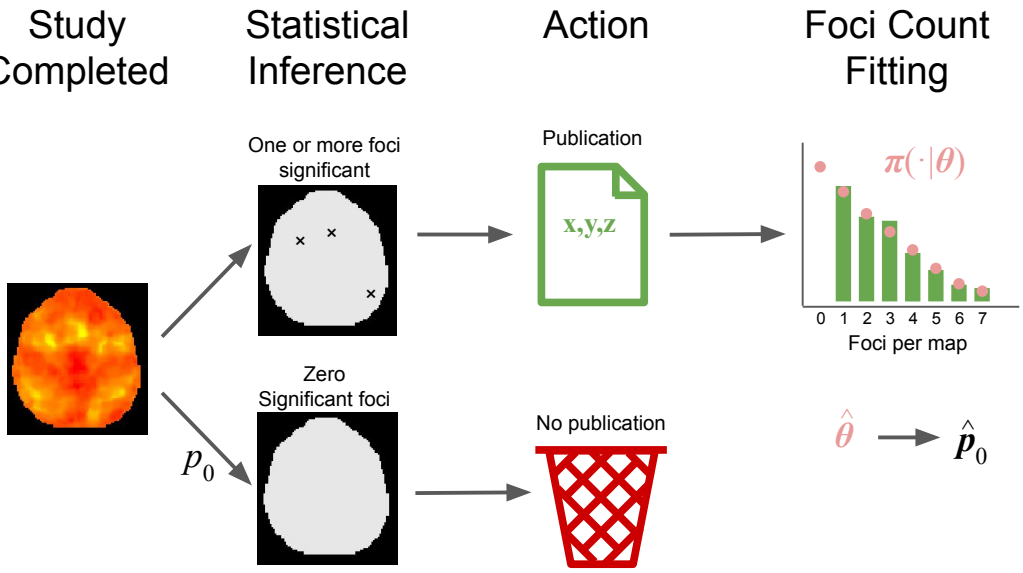

Figure 1: Graphical outline of our approach. We start a population of K experiments, where k of them are published and the remaining $K - k$ are not observed due to lack of significance. Assuming $\pi(n|\boldsymbol{\theta})$ to be the probability mass function of the number of foci per experiment, we use the k published experiment to draw inference on $\boldsymbol{\theta}$ and hence estimate $p_0 = \pi(0 \mid \boldsymbol{\theta})$.

Note that rarely if ever will be able to know the total count of contrasts k, no less K. Hence our approach can be used to learn the *relative* but not the *absolute* frequency of null contrasts. For the latter, it would be necessary to know k, however it is not possible. While our method cannot estimate p_0 in individual meta-analyses, p_0 reflects the sampled population, and thus relates to all studies that make up the population.

Finally, we are careful not to describe our estimators as 'null study' prevalence. Rather, we are estimating prevalence of null contrasts. Each study (paper) consists of multiple contrasts, some of which

might be null. Therefore, since studies typically involve multiple contrasts, we expect that the prevalence of missing studies is much lower compared to the prevalence of missing experiments.

The remainder of the paper is organised as follows. In Section 2, we describe the CBMA data, both real and simulated, that we used and the statistical model for point data that accounts for missing experiments. In Section 3, we present the results of our simulation studies and real data analyses. Finally, in Section 4 we conclude with a discussion of our main findings and set directions for future research.

2 Methods and Materials

2.1 BrainMap database

Our analysis is motivated by coordinate data from *BrainMap*[1] [12, 13, 14, 15]. BrainMap is an online, freely accessible database of coordinate-based data from both functional and structural neuroimaging experiments. The data are exclusively obtained from peer-reviewed papers on whole-brain, voxel-wise studies, that are written in English language.

There are three possible routes via which a paper can enter the database. Firstly, some of the papers are coded by their authors in *Scribe*[2] and are then submitted to the BrainMap team for quality control, either in the process of performing a meta-analysis or subsequently. This accounts for approximately one half of all data. Secondly, some of them are submitted by authors in alternative formats (e.g. spreadsheet) after publication and are then coded into the database by BrainMap staff through Scribe. Thirdly, some papers are retrieved and coded exclusively by BrainMap staff who perform regulars scans of the literature, with a focus on large-scale CBMAs. In these cases, BrainMap staff solicit data from the authors of each paper.

Thanks to these contributions, BrainMap has been continuously expanding since being introduced in 1992. It currently includes three sectors: task activation (TA), voxel-based morphometry (VBM) and voxel-based physiology (VBP). As of April 2019, the TA sector consists of results obtained from 3,502 scientific papers, including both PET and fMRI task-activation data. Each scan condition is coded by stimulus, response, and instruction and experiments are coded most typically as between-condition contrasts. BrainMap TA is publicly available and is the sector which we use in this paper. BrainMap VBM has been recently introduced and contains results from 992 papers (as of April 2019). It consists largely of between-group morphometric contrasts, typically of patients to controls. VBM data include both grey-matter and white-matter contrasts, coded separately. This is sector is publicly available. For more details about BrainMap VBM, see [16]. The VBP sector consists largely of between group physiological contrasts, typically of patients to controls VBP data include cerebral blood flow (PET, SPECT and fMRI), cerebral glucose metabolism (PET), cerebral oxygen metabolism (PET), and indices of neurovascular coupling (fMRI ALFF, ReHO and others). BrainMap VBP is not yet public but can be accessed upon request.

Due to its size, BrainMap is a widely used resource for neuroimaging meta-analysis. More specifically, there are currently (as of April 2019) 861 peer-reviewed articles using the BrainMap and/or its CBMA software. Some recent examples include [17], [18] and [19]. Throughout this paper, we assume that the database is indicative of the population of non-null neuroimaging studies; we discuss the plausibility of this assumption in Section 4.

In this work we do not consider any of the resting-state (because resting-state studies are currently under-represented) studies registered in BrainMap TA. Our unit of observation is a contrast, and hence our dataset consists of 16,285 observations; these are all the contrasts retrieved from the 3,492 papers that we considered. Each observation (contrast) consists of a list of three dimensional coordinates $\boldsymbol{z}_i$, the *foci*, typically either local maxima or centers of mass of voxel clusters with significant activations. For the purposes of this work, we do not use the coordinates, and model the file drawer solely based on the total number of foci per contrast n_i. Table 1 presents some summary statistics for the subset of the BrainMap dataset that we use (*i.e.* functional experiments excluding resting state), whereas Figure 2 shows the empirical distribution of the total number of foci per contrast.

Table 1: BrainMap database summaries.

Database composition

[1] RRID:SCR_003069

[2] Scribe is BrainMap's software to organise each study's results (peak coordinates, associated statistics, significance levels, etc.) and meta-data (study context, sample size, etc.).

Publications	Contrasts	Foci
3,492	16,285	127,713

Contrasts per publication

Min.	Q_1	Median	Mean	Q_3	Max.
1	2	4	4.7	6	63

Foci per contrast

Min.	Q_1	Median	Mean	Q_3	Max.
1	2	5	7.8	11	98

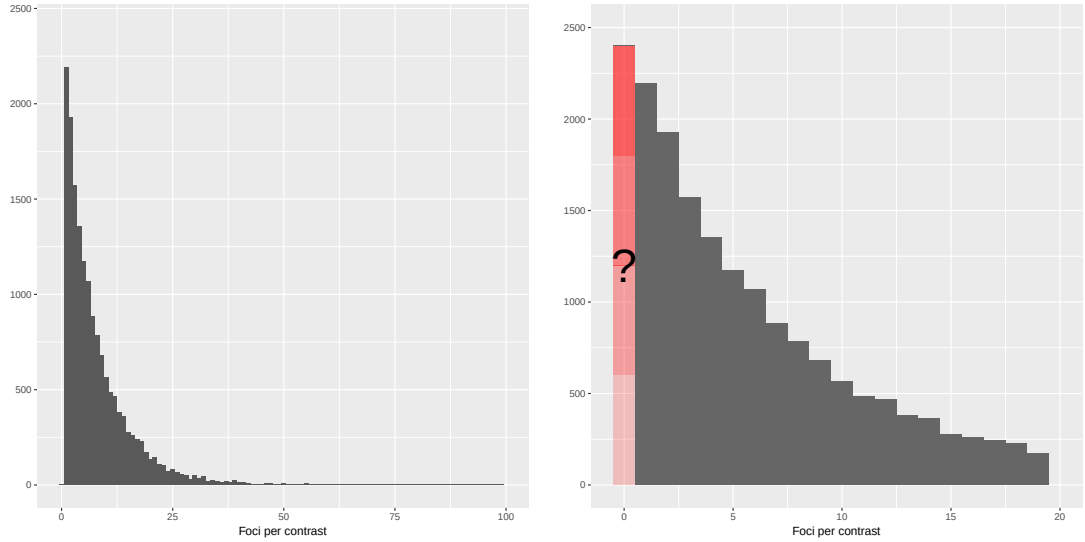

Figure 2: Empirical distribution of the total number of foci per contrast in the BrainMap database, n_i. The left panel shows the full distribution, while the right panel shows a zoomed-in view of all experiments reporting 24 or fewer foci. The BrainMap database does not record incidents of null contrasts (contrasts in a paper for which $n_i = 0$).

The barplot of Figure 2 (right) identifies a fundamental aspect of this data: even though the distribution of n_i has most of its mass close to zero, there are no contrasts with zero foci. This is expected as by design, the contrasts from a paper that report no activations are not registered into the BrainMap database. The objective of this work is to identify the relative proportion of these contrasts compared to the ones that are registered. Some of these null contrasts may in fact be clearly reported in the papers but not registered in the BrainMap database. However, given the stigma of the negative findings, we suspect that they are rare.

2.2 Models

As discussed earlier, our model uses count data from the observed, reported experiments to infer on the file drawer quantity. At this point, we list the two critical assumptions: I) data $\{n_i\}_{i=1}^{k}$, both observed and unobserved, are taken to be independent and identically distributed (i.i.d.) samples from a count distribution N of a given parametric form (we will relax this assumption later, to allow for inter-experiment covariates); II) the probability of publication equals zero for experiments (contrasts) with $n_i = 0$. Assumption II implies that a paper will not appear in BrainMap only if all its contrasts are negative. For a detailed discussion of the implications of assumptions I-II, see Section 4.

As each paper in the BrainMap database has multiple contrasts, potentially violating the independence assumption, we draw subsamples such that exactly one contrast from each publication is used. Specifically, we create 5 subsamples (A-E) drawing 5 different contrasts for each subsample, if possible; for publications with less than 5 contrasts we ensure that every contrast is used in at least one subsample, and then randomly select one for the remaining subsamples.

If assumptions I-II described above hold, then a suitable model for the data is a *zero-truncated* count distribution. A zero-truncated count distribution occurs when we restrict the support of a count

distribution to the positive integers. For a probability mass function (pmf) $\pi(n \mid \boldsymbol{\theta})$ defined on $n = 0, 1, \ldots,$ where $\boldsymbol{\theta}$ is the parameter vector, the zero truncated pmf is:

$$\pi_{\mathrm{ZT}}(n \mid \boldsymbol{\theta}) = \mathbb{P}(N = n) = \frac{\pi(n \mid \boldsymbol{\theta})}{1 - \pi(0 \mid \boldsymbol{\theta})}, \quad n = 1, 2, \ldots. \tag{1}$$

We consider three types of count distributions $\pi(n \mid \boldsymbol{\theta})$: the Poisson, the Negative Binomial and the Delaporte. The Poisson is the classic distribution for counts arising from series of independent events. In particular, if the foci in a set of experiments arise from a spatial Poisson process with common intensity function, then the resulting counts will follow a Poisson distribution. Poisson models often fit count data poorly due to *over-dispersion*, that is, the observed variability of the counts is higher than what would be anticipated by a Poisson distribution. More specifically, if a spatial point process has a random intensity function, one that changes with each experiment, the distribution of counts will show this over-dispersion.

The Negative Binomial distribution is the count distribution arising from the Poisson-Gamma mixture: if the true Poisson rate differs between experiments and is distributed as a Gamma random variable, then the resulting counts will follow a Negative Binomial distribution. For the Negative Binomial distribution we use the mean-dispersion parametrisation:

$$\pi(n \mid \mu, \phi) = \left(\frac{\phi}{\phi + \mu}\right)^{\phi} \frac{\Gamma(\phi + n)}{\Gamma(\phi)} \left(\frac{\mu}{\mu + \phi}\right)^{n}, \tag{2}$$

where μ is the mean, $\phi > 0$ is the dispersion parameter and $\Gamma(\cdot)$ represents the gamma function; with this parametrisation the variance is $\mu + \frac{\mu^2}{\phi}$. Hence, the excess of variability compared to the Poisson model is accounted for through the additional term $\frac{\mu^2}{\phi}$.

The Delaporte distribution is obtained by modelling the foci counts n_i of experiment i as $\mathrm{Pois}(\mu\gamma_i)$ random variables; the γ_i follows a particular shifted Gamma distribution with parameters σ and ν, $\sigma > 0$ and $0 \leq \nu < 1$ [20]. The probability mass function of the Delaporte distribution can be written as:

$$\pi(n \mid \mu, \sigma, \nu) = \frac{\exp(-\mu\nu)}{\Gamma(\frac{1}{\sigma})} \left[1 + \mu\sigma(1 - \nu)\right]^{-\frac{1}{\sigma}} S, \tag{3}$$

where μ is the mean and:

$$S = \sum_{j=0}^{n} \binom{n}{j} \frac{\mu^n \nu^{n-j}}{n!} \left[\mu + \frac{1}{\sigma(1 - \nu)}\right]^{-j} \Gamma\left(\frac{1}{\sigma} + j\right). \tag{4}$$

With this parametrisation the variance of the Delaporte distribution is $\mu + \mu^2\sigma(1 - \nu)^2$.

Once the parameters of the truncated distribution are estimated, one can make statements about the original, untruncated distribution. One possible way to express the file drawer quantity that we are interested in is the percent prevalence of zero count contrasts p_{z}, that is, the total number of missing experiments per 100 published. This can be estimated as:

$$\hat{p}_{\mathrm{z}} = \frac{\pi(0 \mid \hat{\boldsymbol{\theta}})}{1 - \pi(0 \mid \hat{\boldsymbol{\theta}})} \times 100. \tag{5}$$

Here, $\pi(0 \mid \hat{\boldsymbol{\theta}})$ denotes the probability of observing a zero count contrast, and $\hat{\boldsymbol{\theta}}$ denotes the estimated parameter values from the truncated model (e.g. $\boldsymbol{\theta} = (\mu, \sigma, \nu)^{\top}$ for the Delaporte model).

Our statistical model is based on homogeneous data, and we can reasonably expect that differences in experiment type, sample size, etc., can introduce systematic differences between experiments. To explain as much of this nuisance variability as possible, we further model the expected number of foci per experiment as a function of its characteristics in a log-linear regression:

$$\mu = \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right), \tag{6}$$

where $\mathbf{x}_i$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients. The covariates that we consider are: i) the year of publication ranging from 1985 to 2018; ii) the square root of the number of participants[3] ranging from 1 to 395; iii) the experimental context. In each subsample, we merge all the labels of the variable context that are missing or appear less than 20 times into the 'Other'

[3]Since we expect the power to scale with the square root of the sample size.

category. The remaining categories (that appear in at least one subsample) are: aging, disease, disease/emotion, disease/pharmacology, disease/treatment, experimental design/normal mapping, gender, language, learning, normal mapping and pharmacology. Summaries of the BrainMap subsamples A-E data for each level of context can be found in A.

Parameter estimation is done under under the *generalized additive models for location scale and shape* (GAMLSS) framework of [21]. The fitting is done in R[4] [22] with the *gamlss* library [23]. Confidence intervals are obtained with the bootstrap. When covariates are included in the model, we use the stratified bootstrap to ensure representation of all levels of the experimental context variable. In particular, for each level of the categorical variable a bootstrap subsample is drawn using the data available for this class and subsequently these subsamples are merged to provide the bootstrap dataset. Model comparison is done using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) provided by the package.

2.3 Monte Carlo evaluations

We perform a simulation study to assess the quality of estimates of p_z, the total number of experiments missing per 100 published, obtained by the zero-truncated Negative Binomial and Delaporte models (initial work found BrainMap counts completely incompatible with the Poisson model, and hence we did not consider it for simulation). For both approaches, synthetic data are generated as follows. First, we fix the values of the parameters, that is, μ, ϕ for the Negative Binomial distribution and μ, σ, ν for the Delaporte distribution. We then generate $k^*/(1-\pi(0|\boldsymbol{\theta}))$ samples from the untruncated distributions, where k^* is chosen such that the expected number of non-zero counts is k. We remove the zero-count instances from the simulated data and the corresponding zero-truncated model is fit to the remaining observations. Finally, we estimate the probability of observing a zero count experiment based on our parameter estimates.

We set our simulation parameter values to cover typical values found in BrainMap (see C, Table 8). For the Negative Binomial distribution we consider values 4 and 8 for the mean and values 0.4, 0.8, 1.0 and 1.5 for the dispersion, for a total of 8 parameter settings. For the Delaporte distribution, we set μ to 4 and 8, σ to 0.5, 0.9 and 1.2, and ν to 0.02, 0.06 and 0.1 (18 parameter settings). The expected number of observed experiments is set to $k = 200$, 500, 1,000 and 2,000. For each combination of (k, μ, ϕ) and (k, μ, σ, ν) of the Negative Binomial and Delaporte models, respectively, we generate 1,000 datasets from the corresponding model, for each parameter setting, and record the estimated value of p_z for each fitted dataset.

2.4 HCP real data evaluations

As an evaluation of our methods on realistic data for which the exact number of missing contrasts is known, we generate synthetic meta-analysis datasets using the Human Connectome Project task fMRI data. We start with a selection of 80 unrelated subjects and retrieve data for all 86 tasks considered in the experiment. For each task, we randomly split the 80 subjects into 8 groups of 10 subjects. Hence, we obtain a total of $86 \times 8 = 688$ synthetic fMRI experiments. For each experiment, we perform a one-sample group analysis, using ordinary least squares in FSL[5], and recording n_i^{v}, the total number of surviving peaks after random field theory thresholding at the voxel level, 1% familywise error rate (FWE), where $i = 1, \ldots, 688$. We also record the total number of peaks (one peak per cluster) after random field theory thresholding at the cluster level, cluster forming threshold of uncorrected P=0.00001 & 1% FWE, n_i^{c}. These rather stringent significance levels were needed to induce sufficient numbers of results with no activations. We then discard the zero-count instances from n_i^{v} and n_i^{c}, and subsequently analyse the two truncated samples in two separate analyses, using the zero-truncated Negative Binomial and Delaporte models. Finally, the estimated number of missing experiments is compared to the actual number of discarded contrasts. Note that we repeat the procedure described above 6 times, each time using different random splits of the 80 subjects (HCP splits 1-6).

[4] RRID:SCR_001905
[5] RRID:SCR_002823

3 Results

3.1 Simulation results

The percent relative bias of the estimates of p_z, $\frac{\hat{p}_z - p_z}{p_z} \times 100$, and its bootstrap standard error for the zero-truncated Negative Binomial and Delaporte models are shown in Table 2 and Table 3, respectively. The results indicate that, when the model is correctly specified, both approaches perform adequately. In particular, in Table 2 we see that the bias of $\hat{p}_z$ is small, never exceeding 8% when the sample size is comparable to the sample size of the BrainMap database ($k = 3,492$) and the mean number of foci is similar to the average foci count found in BrainMap (≈ 9). The bootstrap standard error estimates produced by the Negative Binomial model are also accurate with relative bias below 5% in most scenarios with more than 500 contrasts, while Delaporte tends to underestimate standard errors but never more than -15% (see Table 3).

Table 2: Percent relative bias for estimation of p_z, the zero-count experiment rate as a percentage of observed experiments, for Negative Binomial and Delaporte models as obtained from 1,000 simulated datasets. Parameter μ is the expected number of foci per experiment, ϕ, σ and ν are additional scale and shape parameters. Negative Binomial performs well and, while Delaporte often underestimated p_z, with at least 1,000 contrasts it always has bias less than 10% (positive bias over-estimates the file drawer problem).

Negative Binomial

Parameter values				% relative bias of $\hat{p}_z$			
μ	ϕ	p_z	$\mathbb{E}[k] =$ **200**	**500**	**1000**	**2000**
4	0.4	62.1	8.76	2.85	1.40	0.80
4	0.8	31.3	2.72	1.97	-0.78	0.32
4	1.0	25.0	1.70	1.21	0.72	0.16
4	1.5	16.6	2.85	1.66	0.71	0.13
8	0.4	42.0	7.07	3.17	0.70	-0.22
8	0.8	17.2	4.35	1.57	0.49	0.21
8	1.0	12.5	3.32	0.84	0.63	0.04
8	1.5	6.7	2.53	0.90	0.69	0.21

Delaporte

Parameter values				% relative bias of $\hat{p}_z$			
μ	σ	ν	p_z	$\mathbb{E}[k] =$ **200**	**500**	**1000**	**2000**
4	0.5	0.02	11.8	-12.65	-10.66	-9.69	-8.02
4	0.9	0.02	20.8	-17.47	-12.35	-10.28	-10.10
4	1.2	0.02	27.6	-20.46	-18.59	-16.64	-13.83
4	0.5	0.06	10.5	-6.13	-5.58	-4.77	-3.73
4	0.9	0.06	18.0	-4.08	-4.18	-1.32	0.15
4	1.2	0.06	23.4	-5.07	-3.27	-1.09	0.01
4	0.5	0.10	9.3	-4.53	-3.32	-2.65	-1.90
4	0.9	0.10	15.6	1.07	3.86	1.91	1.80
4	1.2	0.10	20.0	4.46	3.32	3.89	4.07
8	0.5	0.02	3.6	-13.77	-10.02	-7.75	-5.91
8	0.9	0.02	9.2	-11.99	-9.00	-7.46	-5.04
8	1.2	0.02	13.8	-14.07	-11.22	-9.48	-8.12
8	0.5	0.06	2.8	-3.04	-3.18	-2.12	-1.89
8	0.9	0.06	6.8	1.41	4.13	2.74	2.47
8	1.2	0.06	10.0	10.49	7.52	7.91	5.19
8	0.5	0.10	2.2	0.93	1.36	0.82	0.74
8	0.9	0.10	5.0	8.09	5.88	5.78	4.94
8	1.2	0.10	7.3	17.91	10.68	7.77	4.76

Table 3: Percent relative bias of bootstrap standard error of $\hat{p}_z$, missing experiment rate as a percentage of observed experiments, for Negative Binomial and Delaporte models as obtained from 1,000 simulated datasets. Parameter μ is the expected number of foci per experiment and ϕ, σ and ν are additional scale and shape parameters. For a sample of at least 1,000 contrasts, Negative Binomial standard errors are usually less than 3% in absolute value; while Delaporte has worse bias, it is never less than -15% (negative standard error bias leads to over-confident inferences).

Negative Binomial

Parameter values				% relative bias of $se(\hat{p}_z)$			
μ	ϕ	p_z	$\mathbb{E}[k] = 200$	500	1000	2000
4	0.4	62.1	34.85	8.72	6.26	-1.33
4	0.8	31.3	8.15	-1.08	-1.76	-1.45
4	1.0	25.0	10.04	5.87	1.40	1.10
4	1.5	16.6	3.97	1.20	-0.37	-3.00
8	0.4	42.0	27.65	2.53	2.61	1.31
8	0.8	17.2	4.67	-0.88	0.58	3.29
8	1.0	12.5	1.77	2.76	-0.75	-0.04
8	1.5	6.7	1.43	-0.40	-2.48	-1.32

Delaporte

Parameter values				% relative bias of $se(\hat{p}_z)$			
μ	σ	ν	p_z	$\mathbb{E}[k] = 200$	500	1000	2000
4	0.5	0.02	11.8	-6.98	-6.51	-7.28	-7.23
4	0.9	0.02	20.8	-10.01	-10.28	-11.56	-11.63
4	1.2	0.02	27.6	-5.88	-9.20	-12.08	-11.48
4	0.5	0.06	10.5	-8.80	-5.50	-9.71	-11.00
4	0.9	0.06	18.0	-8.08	-8.87	-13.53	-14.39
4	1.2	0.06	23.4	-2.69	-6.61	-11.58	-13.36
4	0.5	0.10	9.3	-3.09	-3.43	-8.38	-6.59
4	0.9	0.10	15.6	-7.18	-10.05	-8.81	-9.96
4	1.2	0.10	20.0	-10.47	-9.99	-12.13	-13.47
8	0.5	0.02	3.6	-8.74	-6.96	-6.87	-8.49
8	0.9	0.02	9.2	-10.35	-8.84	-8.81	-10.31
8	1.2	0.02	13.8	-2.86	-6.91	-11.28	-13.51
8	0.5	0.06	2.8	-5.93	-9.27	-11.21	-5.80
8	0.9	0.06	6.8	-6.61	-6.70	-10.93	-9.43
8	1.2	0.06	10.0	-1.42	-10.57	-9.72	-10.42
8	0.5	0.10	2.2	-9.40	-8.75	-7.14	-5.62
8	0.9	0.10	5.0	-10.02	-8.39	-8.16	-2.02
8	1.2	0.10	7.3	-8.16	-8.13	-0.26	-0.85

3.2 HCP synthetic data results

Results of the analysis of the HCP synthetic datasets using the zero-truncated Negative Binomial and Delaporte models are summarised in Figure 3 and Table 4. In Figure 3 we plot the empirical count distributions and the fitted probability mass functions for the 12 datasets considered. For datasets obtained with voxelwise thesholding of the image data, we see that the Delaporte distribution provides a better fit compared to the Negative Binomial qualitatively, by AIC in all 6 datasets, and by BIC in five out of six datasets (Table 4). For clusterwise thresholding, there are fewer peaks in general and their distribution is less variable compared to voxelwise thresholding. Both distributions achieve a similar fit. Here, AIC supports the Negative Binomial model in 4 out of 6 datasets and BIC in 5 out of 6 datasets.

Table 4 reports the true number of missing contrasts n_0, along with point estimates $\hat{n}_0$ and the 95% bootstrap intervals obtained by the zero-truncated models. For voxelwise data, the Negative Binomial model overestimates the total number of missing experiments in all 6 datasets as a consequence of the poor fit to the non-zero counts, while the Delaporte model bootstrap intervals include the true value of n_0 in 5 out of 6 datasets, greatly underestimating n_0 in one dataset. For clusterwise counts, the point estimates obtained by the zero-truncated Negative Binomial model are very close to the true values. Notably, n_0 is included within the bootstrap intervals for all 6 datasets. The Delaporte model underestimates the values of n_0 in all 6 datasets, but the bootstrap intervals include n_0 for 4 out of 6

datasets.

Overall, we find that the zero-truncated modeling approach generally provides good estimates of n_0, with the Negative Binomial sometimes overestimating and the Delaporte sometimes underestimating n_0. A conservative approach, therefore, favors the Delaporte model.

Table 4: Evaluation of the zero-truncated modelling approach using synthetic data obtained from the HCP project, using voxelwise (top) and clusterwise (bottom) inference. The true number of missing contrasts (n_0) for each one of the 12 datasets (6 for voxelwise thesholding and 6 for clusterwise thresholding) is shown in the second column. For each of the Negative Binomial and Delaporte methods, the estimated missing contrast count ($\hat{n}_0$), 95% bootstrap confidence interval for n_0, AIC score and BIC score are shown (smaller AIC and BIC are better).

Voxelwise

		Negative Binomial			Delaporte		
Split	$\mathbf{n_0}$	$\hat{\mathbf{n}}_\mathbf{0}$	**AIC**	**BIC**	$\hat{\mathbf{n}}_\mathbf{0}$	**AIC**	**BIC**
1	7	20 [14,27]	6576.4	6585.4	1 [1,4]	6562.8	6576.4
2	5	22 [15,29]	6583.3	6592.4	3 [1,8]	6576.2	6589.8
3	5	22 [16,30]	6575.9	6585.0	1 [1,21]	6566.1	6579.7
4	4	25 [18,33]	6603.8	6612.9	4 [1,25]	6601.7	6615.3
5	10	18 [13,24]	6539.4	6548.5	1 [0,1]	6504.1	6517.6
6	10	21 [15,29]	6557.0	6566.0	2 [1,10]	6550.0	6563.5

Clusterwise

		Negative Binomial			Delaporte		
Split	$\mathbf{n_0}$	$\hat{\mathbf{n}}_\mathbf{0}$	**AIC**	**BIC**	$\hat{\mathbf{n}}_\mathbf{0}$	**AIC**	**BIC**
1	148	167 [115,248]	3167.8	3176.4	71 [50,108]	3166.9	3179.8
2	144	151 [104,217]	3209.3	3217.9	58 [44,79]	3204.5	3217.4
3	150	161 [109,246]	3163.2	3171.8	95 [62,184]	3164.4	3177.2
4	151	154 [107,231]	3156.3	3164.8	106 [57,159]	3158.0	3170.9
5	153	148 [101,291]	3175.0	3183.6	98 [60,174]	3176.5	3189.4
6	152	151 [103,223]	3198.2	3206.8	89 [54,157]	3199.7	3212.5

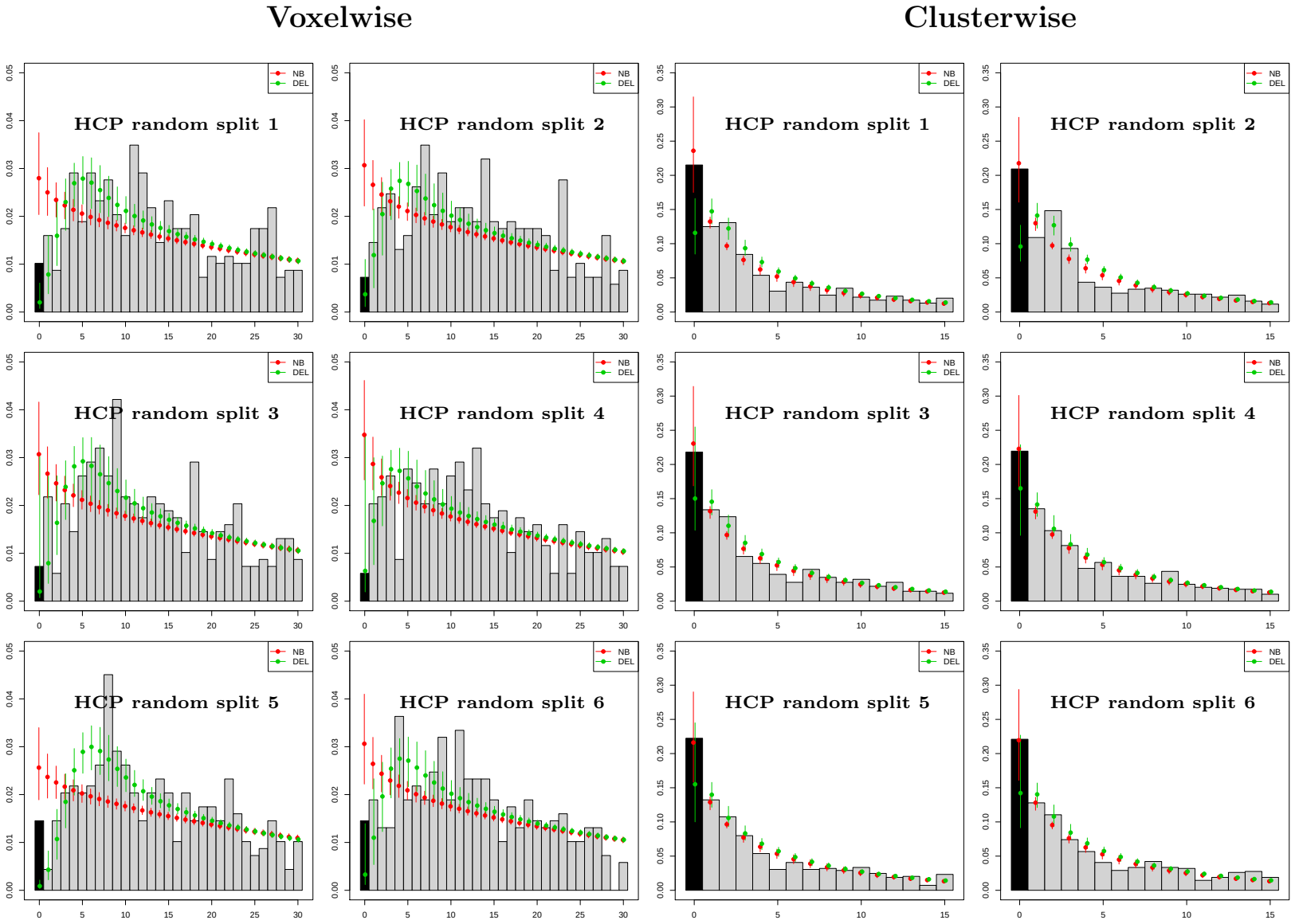

Figure 3: Evaluation with HCP data with 688 contrasts of sample size 10, comparing accuracy of Negative Binomial (NB) and Delaporte (DEL) distributions for the prediction of the number of contrasts with no significant results (zero foci) based on only significant results (one or more foci). Left panel shows results for voxelwise inference, right for clusterwise inference, both using P_{FWE}=0.01 to increase frequency of zero foci. For clusterwise datasets, the Negative Binomial confidence intervals always include the observed zero count, while Delaporte ofter underestimates the count. For voxelwise analysis, the Negative Binomial over-estimates the zero frequency substantially, while Delaporte's intervals include the actual zero frequency in 3 out of 5 splits.

3.3 Application to the BrainMap data

We found the Poisson distribution to be completely incompatible with the BrainMap count data (B, Figure 8), and we do not consider it further. We start by fitting the Negative Binomial and Delaporte zero-truncated models without any covariates. The estimates of the scalar parameters obtained for both models are shown in Appendix C. Figure 4 shows the emprical and fitted probability mass functions for the 5 subsamples. We see that both distributions provide a good fit for the BrainMap data. The Negative Binomial model is preferred based on AIC in 4 out of 5 subsamples, and based on BIC in 5 out of 5 subsamples (Table 6), buth with little difference in both criteria. The estimated prevalence of missing contrasts, along with 95% bootstrap intervals are shown in Table 5. Note that while there is considerable variation in the estimates over the two models, the confidence intervals from all subsamples do not include zero, thus suggesting a file drawer effect.

Table 5: BrainMap data analysis results. The table presents the estimated prevelance of file drawer experiments along with 95% bootstrap confidence intervals, as obtained by fitting the zero-truncated Negative Binomial and Delaporte models to BrainMap subsmaples A-E. No covariates are considered.

	Negative Binomial		**Delaporte**	
Subsample	$\hat{\mathbf{p}}_z$	**95% interval**	$\hat{\mathbf{p}}_z$	**95% interval**
A	11.80	[10.22,13.59]	10.52	[6.42,13.04]
B	11.14	[9.63,12.79]	7.30	[5.29,11.44]
C	11.46	[9.95,13.31]	8.30	[5.95,12.04]
D	11.16	[9.64,12.88]	6.83	[5.03,10.97]
E	11.40	[9.89,13.14]	8.11	[5.44,12.30]

For both Negative Binomial and Delaporte models, and all subsamples A-E, the model with covariates is preferred over the simple model (without the covariates) in terms of AIC but not in terms of BIC (Table 6). This is expected since BIC penalizes model complexity more heavily. Covariates essentially have no effect on the estimated prevalence of missing contrasts. As can be seen in Figure 5, the estimated prevalence of zero count contrasts is a slowly decreasing function of both the square root number of participants and the year of publication. For the former, the trend is expected and one possible explanation is that bigger samples result into greater power, and therefore more foci and thus less of a file drawer problem. However, for publication year, decreasing publication bias is welcomed but we could have just as well expected that the increased use of multiple testing in later years would have reduced foci counts and *increased* the file drawer problem. We further see that the estimated percent prevalence of zero-count contrasts is similar for all levels of the categorical variable context, with the exception of experiments studying gender effects (Figure 6). Finally, when including the covariates, the Negative Binomial is preferred over the Delaporte in 3 out of 5 subsamples in terms of the AIC, and in 4 out of 5 subsamples in terms of the BIC (Table 6).

Table 6: AIC/BIC model comparison results for the BrainMap data. We fit the zero-truncated Negative Binomial and Delaporte models, with and without the covariates, to BrainMap subsamples A-E. Every split indicates evidence for better fit with the Negative Binomial model (smaller AIC/BIC indicates better fitting model). For regression models, the sample size was smaller due to missing values. Hence, the criteria cannot be used to compare the models without and with the covariates.

Model comparison: no covariates

	Negative Binomial		**Delaporte**	
Subsample	**AIC**	**BIC**	**AIC**	**BIC**
A	21837.87	21850.18	21839.87	21858.34
B	21792.04	21804.36	21792.61	21811.09
C	21899.71	21912.03	21901.18	21919.65
D	21862.23	21874.54	21861.83	21880.30
E	21683.47	21695.79	21684.94	21703.42

Model comparison: regression

	Negative Binomial		**Delaporte**	
Subsample	**AIC**	**BIC**	**AIC**	**BIC**

A	21778.73	21871.10	21780.45	21878.98
B	21735.59	21827.96	21733.23	21831.76
C	21850.23	21936.44	21850.29	21942.66
D	21816.27	21902.49	21813.07	21905.45
E	21628.24	21714.45	21628.58	21720.95

4 Discussion

4.1 Summary of findings & implications for CBMA

In this paper, we have attempted to estimate the prevalence of experiments missing from a CBMA due to reporting non-significant results. Our method uses intrinsic statistical characteristics of the non-zero count data to infer the relative frequency of zero counts. This is achieved by estimating the parameters of a zero-truncated model, either Negative Binomial or Delaporte, which are subsequently used to predict the prevalence p_0 of zero-count experiments in the original, untruncated distribution, and re-expressing this as p_z, the rate of missing contrasts per 100 published.

Our approach further relies on assumptions I and II described in Section 2.2. Assumption I implies that there is independence between contrasts. However, as one publication can have several contrasts, this assumption is tenuous despite it being a standard assumption for most CBMA methods. To ensure the independence assumption is valid, we subsample the data so that only one randomly selected contrast per publication is used. Assumption II defines our censoring mechanism, such that only experiments with at least one significant activation can be published. The assumption that non-significant research findings are suppressed from the literature has been adopted by authors in classical meta-analysis [24, among others]. One possible way in which this assumption can be violated could be due to data repeatedly analysed under different pipelines (e.g. by using a different cluster extent each time) until they provide significant results. However, we believe that researcheres are unlikely to resort to this approach because studies typically involve multiple contrasts. Hence, even if some of them are negative, the authors can focus on remaining, significant contrasts, in their publication. Assumption II can also be violated due to contrasts that have significant findings but are not reported because these findings are not in agreement with the researcher's hypothesis of interest or the existing literature. However, this violation is not an issue unless the distribution of n in such contrasts is different to the distribution of n in contrasts that are reported.

A series of simulations studies suggest that the zero-truncated modelling approach provides valid estimates of p_z. A critical limitation of our HCP evaluation is the repeated measures structure, where 86 contrasts come from each subject. Such dependence generally does not induce bias in the mean estimates, but can corrupt standard errors and is a violation of the bootstrap's independence assumption. However, as the bootstrap intervals generally captured the true censoring rate, it seems we were not adversely affected by this violation. It should be noted, moreover, that the properties of our estimators degrade as the total number of observed experiments decreases and therefore our methods are likely not suitable for individual meta-analyses unless hundreds of experiments are available.

The analysis of BrainMap data suggests that the estimated prevalence of null contrasts slightly varies depending on the characteristics of an experiment, but generally consists of at least 6 missing experiments for 100 published, and this estimate of 6 is significantly greater than zero. In other words, for a randomly selected CBMA consisting of J contrasts, we expect that $6J/100$ null contrasts are missing due to the file drawer. Note that this interpretation concerns the aggregate statistical practice reflected in BrainMap, i.e. it is totally agnostic to the statistical procedures used to generate the results in the database. The counts we model could have been found with liberal $P < 0.001$ uncorrected inferences or stringent $P < 0.05$ FWE procedures. However, if the neuroimaging community *never* used multiple testing corrections, then every experiment should report many peaks, and we should estimate virtually no missing contrasts.

The results suggest that the population sampled by the BrainMap database has a non-zero file drawer effect. Whether this conclusion can be extended to all neuroimaging experiments depends on the representativeness of the database. As noted above, the BrainMap staff are continually adding studies and capture the content of newly published meta-analyses. Hence, the most notable bias could be topicality and novelty effects that drive researchers to create meta-analyses. Another potential source of bias could be due to studies which BrainMap does not have access to, such as ones that are never published due insufficient time to submit a paper or staff leaving. But we do not see these particular

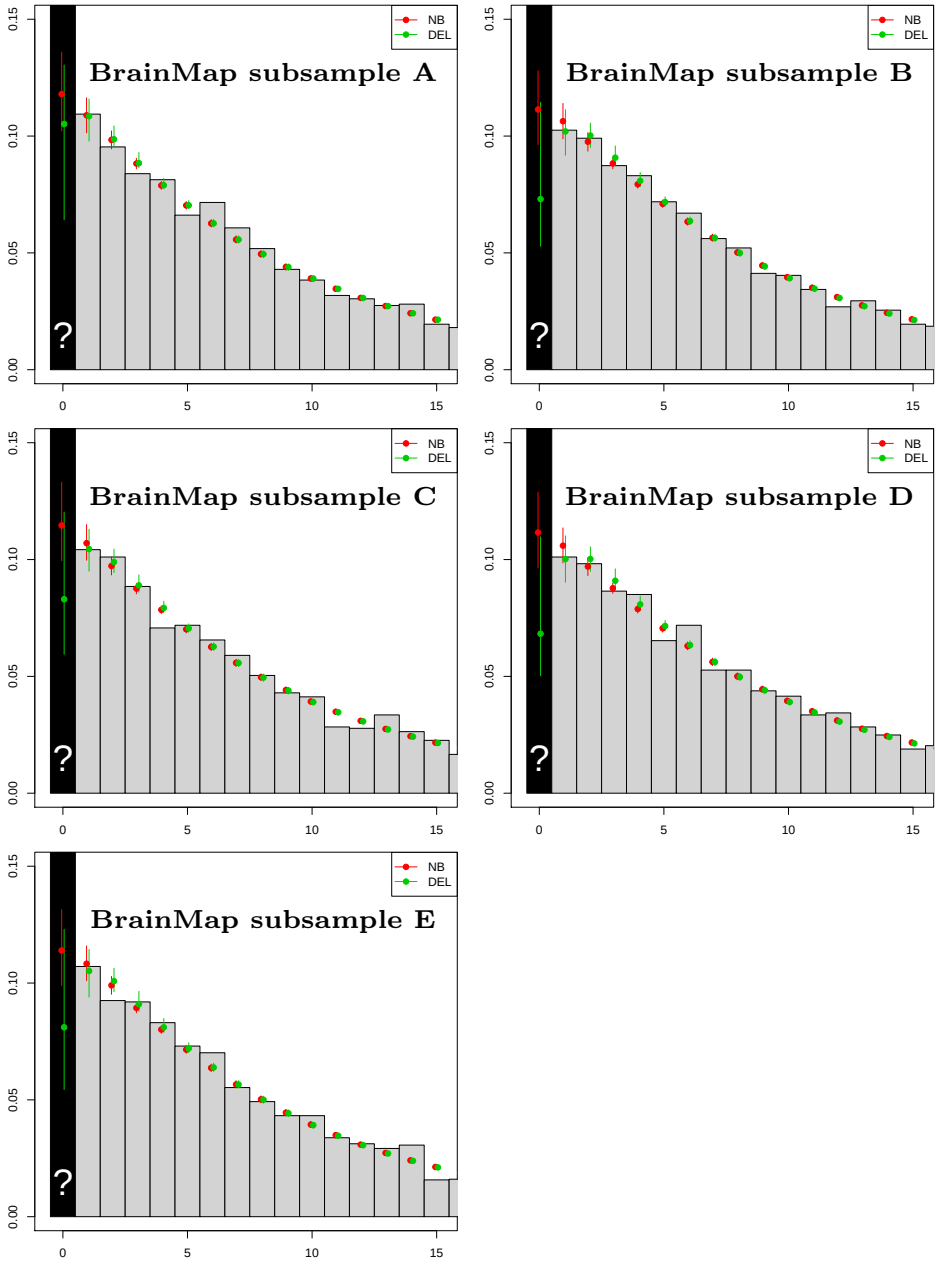

Figure 4: BrainMap results for 5 random samples using the Negative Binomial and Delaporte models and no covariates. Plots show observed count data (gray bars) with fit of full (non-truncated) distribution based on zero-truncated data, including the estimate of p_0 (over black bar).

effects as driving the file drawer effect up or down in particular, and so is not so much of a concern.

Our findings provide evidence for the existence of publication bias in CBMA. The presence of missing experiments does not invalidate existing CBMA studies, but complements the picture seen when conducting a literature review. Considering the missing contrasts would affect the conclusions drawn from any of the existing CBMA approaches. For model-based approaches based on spatial point processes [25, 26, among others], the inclusion of null contrasts would cause the estimated intensity functions at each voxel to shift downwards, thus leading to potentially different inferences. For kernel-based approaches (such as MKDA [27] ALE [28] and SDM [29], inclusion of null contrasts would also lead to lower values of the estimated statistic at each voxel. However, it would not affect the inferences (i.e. significant voxels) obtained. This is due to the fact that kernel-based approaches are developed in order to test spatial

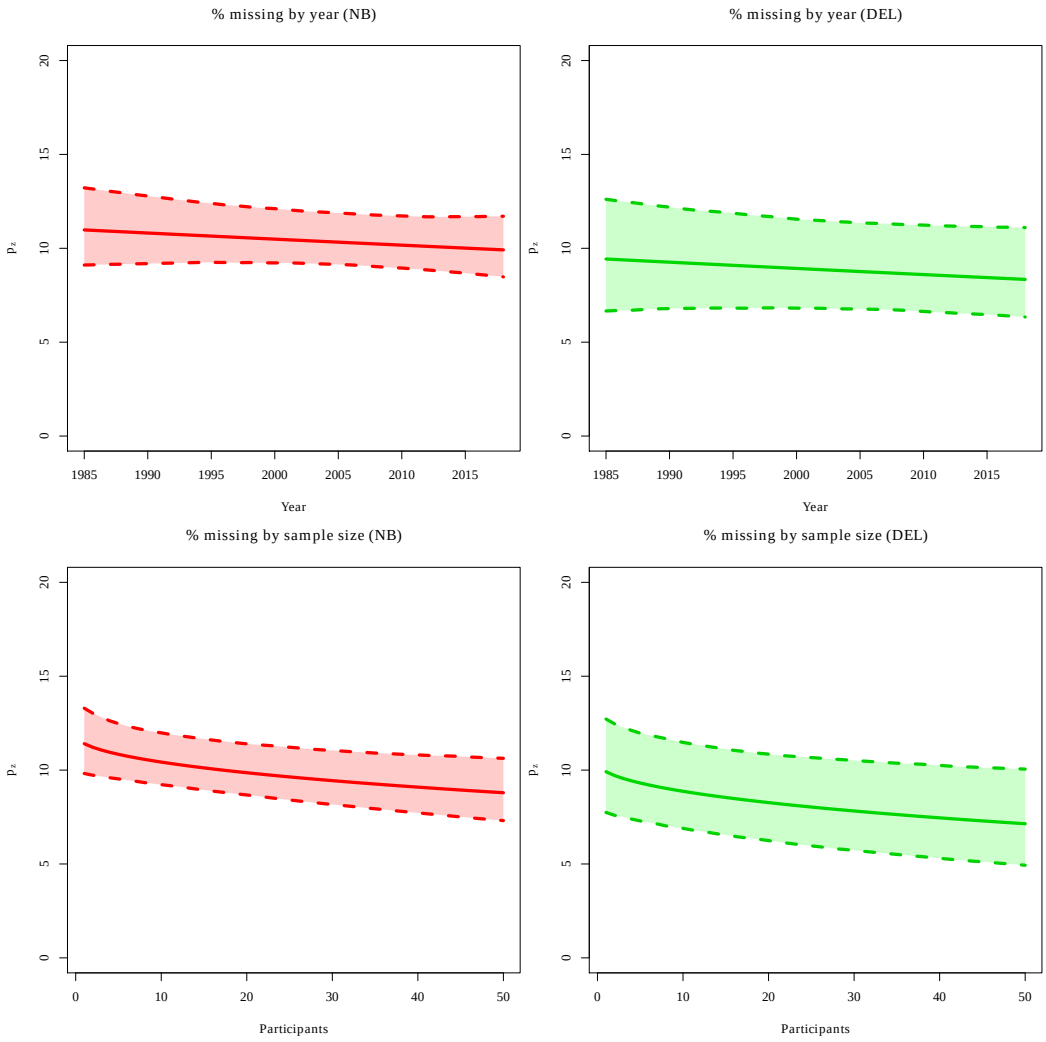

Figure 5: Predicted p_z, missing experiment rate per 100 published experiments, as a function of year of publication (top) and the square root of sample size (bottom), with pointwise 95% bootstrap confidence intervals. There is not much variation in the estimate of the percentage missing, but in both cases a negative slope is observed, as might be expected with improving research practices over time and greater power with increased sample size. All panels refer to the second BrainMap random sample (subsample B).

convergence of reported foci conditional on at least one activation (rather than assessing the evidence for the existence of a population effect).

4.2 Future work

There are a few limitations to our work. Even though we posit that the majority of the missing contrasts are never described in publications or not published at all, we cannot rule out the contribution from contrasts that have actually been reported in the original publications and simply not encoded in BrainMap. Therefore, it is worth considering an extensive literature review in order to investigate how often such null results are mentioned in papers. This information can be then used to approximate the fraction of contrasts that are never published. Ideally, our unit of inference would be a publication rather than a contrast. However, linking our contrast-level inferences to studies requires assumptions about dependence of contrasts within a study and the distribution of the number of contrasts examined per study. We can assert that the more contrasts examined per investigation, the more likely 1 or more null contrasts should arise; and that the risk of null contrasts is inversely related to foci-per-contrast

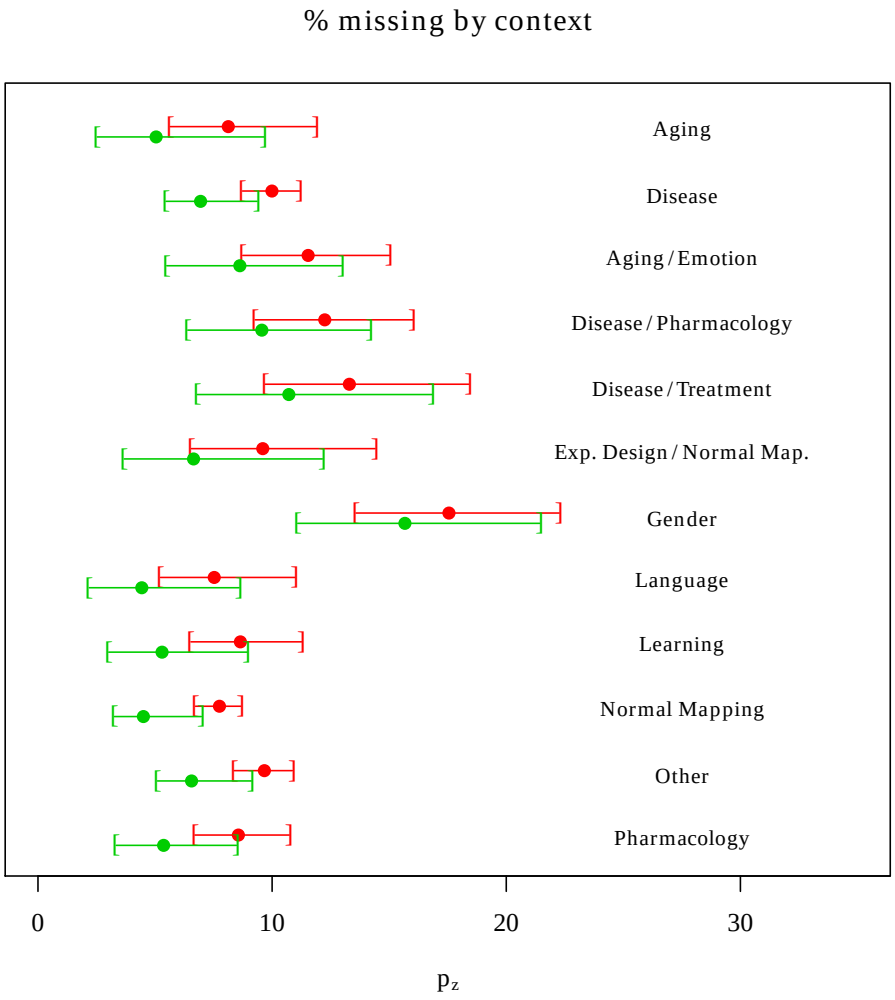

Figure 6: Contrasts missing per 100 published as a function of experiment context, with 95% bootstrap confidence intervals. Note that we have fixed the year and square root sample size covariates to their median values. The plot is derived from the second BrainMap random sample (subsample B). The red and green lines represent the Negative Binomial and Delaporte distributions, respectively.

of non-null contrasts. Using these facts, there might be ways in which one could use our findings to estimate the rate of missing studies.

The evaluation of our methods using the HCP data could be also extended. One option would be to implement a different analysis pipeline in each one of the synthetic experiments in order to reflect the heterogeneity observed in the BrainMap data. Another option would be to investigate the robustness of the results to the choice of the sample size of each synthetic experiment. Nonetheless, for larger sample sizes, this would require a larger selection of HCP subjects in order to ensure that there total number of synthetic experiments is sufficient for our approach. The analysis conducted in this paper is based on data retrieved from a single database. As a consequence, results are not robust to possible biases in the way publications are included in this particular database. A more thorough analysis would require consideration of other databases (e.g. Neurosynth.org[6] [30], though note Neurosynth does not report foci per contrast but per paper).

One may argue that our censoring mechanism is rather simplistic, and does not reflect the complexity of current (and potentially) poor scientific practice. As discussed earlier, we have not allowed for the possibility of 'vibration effects', that is, changing the analysis pipeline (e.g., random vs fixed effects, linear vs. nonlinear registration) to finally obtain some significant activations. This would be an instance

[6]RRID:SCR_006798

of initially-censored (zero-count) data being 'promoted' to a non-zero count through some means, see Figure 7 for a graphical representation. Such models can be fit under the Bayesian paradigm and we will consider them in our future work. Our simulation studies have shown that the properties of our prevalence estimator are poor when the total number of non-zero experiments available is low. This fact implies that the estimator cannot be used to infer the number of missing experiments from a single CBMA. Hence, a potential direction for future work would be to construct estimators that are more robust when the sample size is small.

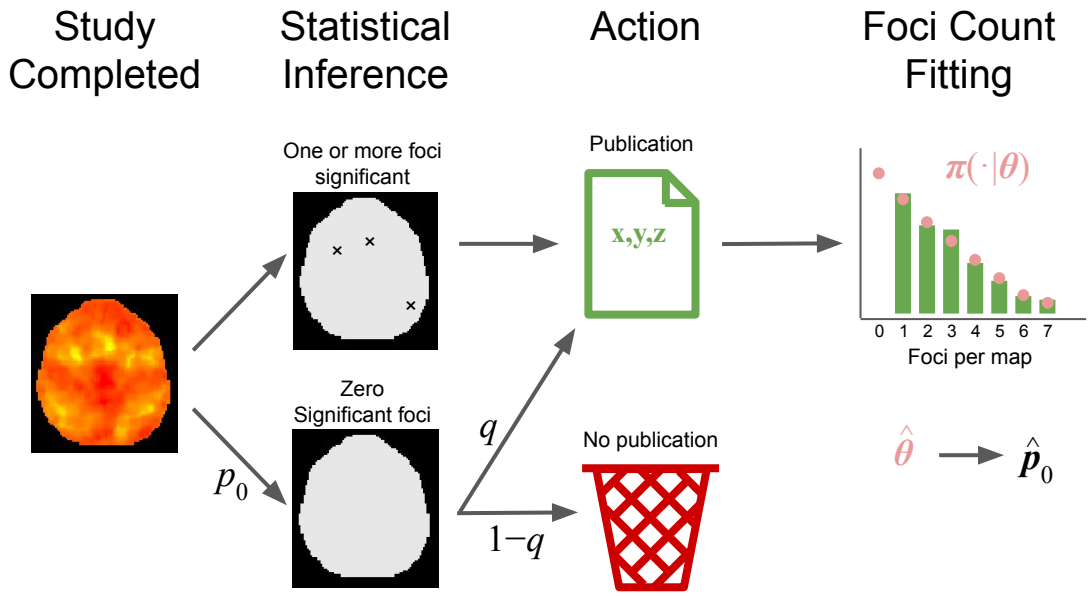

Figure 7: A zero-upgrade model.

Given results in this paper, there are potential benefits in extending existing CBMA methodologies to account for the file drawer. Many authors have suggested possible solutions for this problem in classic meta-analysis, using, for example *funnel plots* [31, 32], *weight functions* [33, 34] or *sensitivity analyses* [35]. For a recent survey of such methods, see [36]. It is therefore conceivable to adapt these methods for application in CBMA. Moreover, it is worth extending CBMA methods based on spatial point process to be zero-truncated. This would mitigate the bias in the estimated intensity functions caused by non-truncated likelihoods being used when the data are, in fact, zero-truncated. Note that such an extension is not required for kernel-based methods as these are conditional on at least one activation. However, researchers should be wary of interpreting the statistics obtained by kernel-based methods as population effects given that it is likely that there are null studies that are not included in their CBMA.

Finally, it is essential to investigate the existence of forms of publication bias other than null file drawer contrasts, such as studies that are not reported due to results conflicting with literature or studies not reported in academic papers. [37] develop a robustness check for ALE method which is based on the *fail-safe N* [6, 7] that is, the minimum number of unpublished studies required to overturn the outcome of meta-analysis. However, it is essential that such checks are developed for other widely-used CBMA methods.

Highlights

What is already known: Coordinate-based neuroimaging meta-analyses, like classic meta-analyses, are subject to the file drawer problem which can substantially bias the estimates obtained.
What is new: Using the BrainMap data, we estimate that the prevalence of missing experiments in coordinate-based meta-analyses is at least 6 experiments per 100 published.
Potential impact: Our results highlight the need for careful interpretation of the findings obtained from a coordinate-based meta-analyses, as well as the need to extend existing methodologies to account for the file drawer problem.

Acknowledgements

The authors are grateful to Daniel Simpson, Paul Kirk and Solon Karapanagiotis for helpful discussions. This work was largely completed while PS, SM and TEN were at the University of Warwick, Department of Statistics. PS, TDJ and TEN were supported by NIH grant 5-R01-NS-075066; TEN was supported by a Wellcome Trust fellowship 100309/Z/12/Z and NIH grant R01 2R01EB015611-04. The study was further supported by NIH grant MH074457. The work presented in this paper represents the views of the authors and not necessarily those of the NIH or the Wellcome Trust Foundation.

Data Availability Statement

The code that we used has been made publicly available at https://osf.io/ayhfv/.

References

[1] Martha J Farah. Brain images, babies, and bathwater: critiquing critiques of functional neuroimaging. *The Hastings Center Report*, 44(S2):S19–S30, 2014.

[2] M. Raemaekers, M. Vink, B. Zandbelt, R.J.A. van Wezel, R.S. Kahn, and N.F. Ramsey. Test-retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, 36(3):532–542, 2007.

[3] Tor D. Wager, Martin A. Lindquist, Thomas E. Nichols, Hedy Kober, and Jared X. Van Snellenberg. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, 45(Supplement 1):S210—-S221, 2009.

[4] Joshua Carp. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289–300, 2012.

[5] S. T. Normand. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999.

[6] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979.

[7] S. Iyengar and J. B. Greenhouse. Selection models and the file drawer problem. *Statistical Science*, 3(1):133–135, 1988.

[8] C. B. Begg and J. A. Berlin. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 151(3):419–463, 1988.

[9] A. J. Sutton, S. J. Duval, R. L. Tweedie, K. R. Abrams, and D. R. Jones. Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, 320(7249):1574–1577, 2000.

[10] Robin G. Jennings and John D. Van Horn. Publication bias in neuroimaging research: Implications for meta-analyses. *Neuroinformatics*, 10(1):67–80, 2012.

[11] S. P. David, J. J. Ware, I. M. Chu, P. D. Loftus, P. Fusar-Poli, J. Radua, M. R. Munafò, and J. P. A. Ioannidis. Potential reporting bias in fMRI studies of the brain. *PLoS ONE*, 8(7):e70104, 2013.

[12] Peter T Fox and Jack L Lancaster. Neuroscience on the net. *Science*, 266(5187):994–996, 1994.

[13] Peter T Fox and Jack L Lancaster. Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319–321, 2002.

[14] Angela R. Laird, P. Mickle Fox, Cathy J. Price, David C. Glahn, Angela M. Uecker, Jack L. Lancaster, Peter E. Turkeltaub, Peter Kochunov, and Peter T. Fox. ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1):155–164, 2005.

[15] Peter T Fox, Jack L Lancaster, Angela R Laird, and Simon B Eickhoff. Meta-analysis in human neuroimaging: computational modeling of large-scale databases. *Annual review of neuroscience*, 37:409–434, 2014.

[16] Thomas J Vanasse, P Mickle Fox, Daniel S Barron, Michaela Robertson, Simon B Eickhoff, Jack L Lancaster, and Peter T Fox. Brainmap vbm: An environment for structural meta-analysis. *Human brain mapping*, 39(8):3308–3325, 2018.

[17] A. C. Hill, A. R. Laird, and J. L. Robinson. Gender differences in working memory networks: A brainmap meta-analysis. *Biological Psychology*, 102(0):18–29, 2014.

[18] Lauren AJ Kirby and Jennifer L Robinson. Affective mapping: An activation likelihood estimation (ALE) meta-analysis. *Brain and Cognition*, 118:137–148, 2017.

[19] Yuwen Hung, Schuyler L Gaillard, Pavel Yarmak, and Marie Arsalidou. Dissociations of cognitive inhibition, response inhibition, and emotional interference: Voxelwise ALE meta-analyses of fMRI studies. *Human Brain Mapping*, 2018.

[20] R.A. Rigby, D.M. Stasinopoulos, and C. Akantziliotou. A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution. *Computational Statistics & Data Analysis*, 53(2):381 – 393, 2008.

[21] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 54(3):507–554, 2005.

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[23] D. M. Stasinopoulos and R. A. Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 12 2007.

[24] Lynn E Eberly and George Casella. Bayesian estimation of the number of unseen studies in a meta-analysis. *Official Journal of Statistics*, 15(4):477–494, 1999.

[25] Jian Kang, Thomas E. Nichols, Tor D. Wager, and Timothy D. Johnson. A Bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *The Annals of Applied Statistics*, 8(3):1561–1582, 2014.

[26] Silvia Montagna, Tor Wager, Lisa Feldman Barrett, Timothy D Johnson, and Thomas E Nichols. Spatial bayesian latent factor regression modeling of coordinate-based meta-analysis data. *Biometrics*, 74(1):342–353, 2018.

[27] Tor D. Wager, Martin Lindquist, and Lauren Kaplan. Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2):150–158, 2007.

[28] Simon B. Eickhoff, Angela R. Laird, Christian Grefkes, Ling E. Wang, Karl Zilles, and Peter T. Fox. 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, 2009.

[29] Joaquim Radua and David Mataix-Cols. Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *The British Journal of Psychiatry : the Journal of Mental Science*, 195(5):393–402, November 2009.

[30] Tal Yarkoni, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–670, 2011.

[31] M. Egger, G. Davey Smith, M. Schneider, and C. Minder. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634, 1997.

[32] S. Duval and R. Tweedie. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.

[33] D. T. Larose and D. K. Dey. Modeling publication bias using weighted distributions in a Bayesian framework. *Computational Statistics and Data Analysis*, 26(3):279–302, 1998.

[34] J. Copas and D. Jackson. A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, 60(1):146–153, 2004.

[35] J. Copas and J. Q. Shi. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1(3):247–262, 2000.

[36] Z. Jin, X. Zhou, and J. He. Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, 34(2):343–360, 2015.

[37] Freya Acar, Ruth Seurinck, Simon B Eickhoff, and Beatrijs Moerkerke. Assessing robustness against potential publication bias in Activation Likelihood Estimation (ALE) meta-analyses for fMRI. *PloS one*, 13(11):e0208177, 2018.

A BrainMap summaries for context

In this section we provide summaries of the data on the 5 BrainMap subsamples A-E, for the different levels of the categorical variable experiment context. In particular, Table 7 presents the total number of contrasts per level, the average sample size per contrast, and the average number of foci per contrast. Note that in subsamples C, D and E there were less than 20 contrasts with label 'Gender'; hence, we incorporate those in the 'Other' category.

Table 7: Data summaries for the different levels of the categorical variable experiment context.

Contrasts per level

Experiment context	BrainMap subsample				
	A	**B**	**C**	**D**	**E**
Aging	20	27	22	23	21
Disease	574	595	593	590	585
Disease, Emotion	34	32	31	33	31
Disease, Pharmacology	48	49	42	47	38
Disease, Treatment	33	30	31	33	28
Experimental Design, Normal Mapping	30	30	32	33	26
Gender	22	23	-	-	-
Language	29	32	29	27	31
Learning	26	31	25	27	28
Normal Mapping	2074	2043	2063	2062	2081
Other	539	542	563	555	566
Pharmacology	63	58	61	62	57

Average constrast sample size

Experiment context	BrainMap subsample				
	A	**B**	**C**	**D**	**E**
Aging	16.4	12.4	12.0	11.7	12.2
Disease	14.4	15.0	14.5	15.1	15.0
Disease, Emotion	15.7	16.1	16.2	16.0	16.0
Disease, Pharmacology	12.3	12.5	12.8	12.8	12.1
Disease, Treatment	11.4	12.9	12.4	11.9	11.3
Experimental Design, Normal Mapping	19.2	19.6	21.7	20.9	20.5
Gender	13.2	12.2	-	-	-
Language	11.6	11.8	11.6	11.9	10.7
Learning	10.4	11.6	10.2	10.5	11.8
Normal Mapping	13.9	13.7	13.8	13.7	13.9
Other	16.9	17.1	16.8	16.9	16.9
Pharmacology	13.0	13.4	12.5	13.2	12.6

Average foci per contrast

Experiment context	BrainMap subsample				
	A	**B**	**C**	**D**	**E**
Aging	10.4	9.1	8.5	9.2	6.0
Disease	7.2	7.5	7.6	7.5	7.3
Disease, Emotion	6.6	6.6	6.5	7.8	6.0
Disease, Pharmacology	5.6	6.0	6.0	6.6	5.2
Disease, Treatment	8.2	5.6	6.5	6.7	7.3
Experimental Design, Normal Mapping	7.6	8.0	7.2	5.9	8.0
Gender	5.8	4.0	-	-	-
Language	9.1	9.9	8.1	7.1	7.9
Learning	7.6	8.5	7.6	7.0	9.6
Normal Mapping	9.7	9.7	9.9	9.7	9.5
Other	8.1	7.9	7.8	8.2	7.9
Pharmacology	8.4	8.8	9.5	8.4	7.7

B Zero-truncated Poisson analysis of the BrainMap dataset

In this section, we present results of the analysis of BrainMap subsamples A-E using the zero-truncated Poisson model. The empirical and fitted Poisson probability mass functions are shown in Figure 8. It is evident that the zero-truncated Poisson model provides a poor fit to the BrainMap data. The finding is confirmed by the AIC and BIC criteria. The AIC is 35513.5, 34886.8, 35595.9, 35456.7 and 34642.1 for subsamples A-E, respectively. The BIC is 35519.7, 34893.0, 35602.1, 35462.8 and 34648.3 for subsamples A-E, respectively. These values are much higher than the corresponding values obtained by fitting both the Negative Binomial and Delaporte models (see Table 6). The estimated prevalence of file drawer experiments is estimated as almost zero in all subsamples (Figure 8, final plot). However, these estimates should not be trusted considering the poor fit provided by the zero-truncated Poisson model.

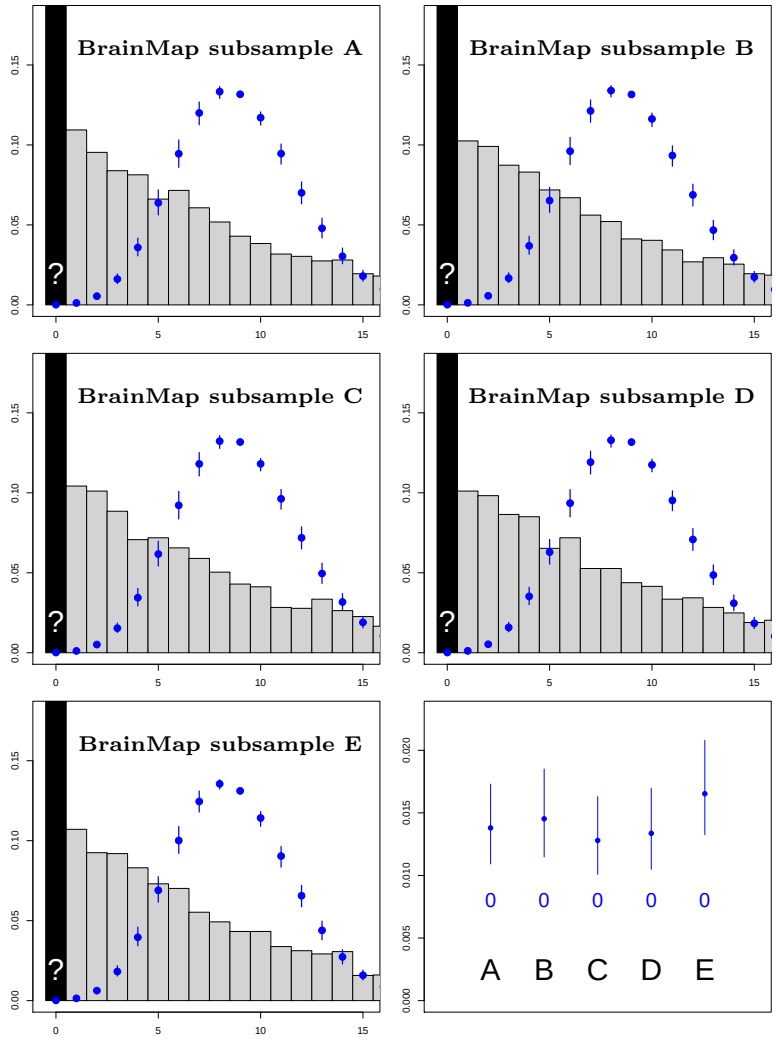

Figure 8: BrainMap results for 5 random samples using the zero-truncated Poisson distribution. The first 5 plots show observed count data (gray bars) with fit of full (non-truncated) distribution based on zero-truncated data, including the estimate of p_0 (over black bar). Final plot shows estimates of p_z, prevalence of file drawer experiments for every 100 experiments observed. All fitted values include 95% bootstrap confidence intervals. The Poisson model provides a poor fit to all 5 subsamples.

C Negative Binomial and Delaporte parameter estimates

In this section, we present the parameter estimates obtained from the analysis of BrainMap subsamples A-E with the simple (without covariates) zero-truncated Negative Binomial and Delaporte models. The parameter estimates are listed in Table 8.

Table 8: Scalar parameter estimates obtained when fitting the simple zero-truncated Negative Binomial and Delaporte models to BrainMap subsamples A-E.

	Negative Binomial		**Delaporte**		
Subsample	$\boldsymbol{\mu}$	$\boldsymbol{\phi}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\nu}$
A	7.95	0.96	8.05	0.96	0.014
B	7.95	0.92	8.24	0.98	0.060
C	8.04	0.95	8.28	0.98	0.043
D	8.03	0.93	8.35	1.01	0.070
E	7.82	0.92	8.06	0.96	0.048