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

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

[1] *MRC Biostatistics Unit, University of Cambridge* [2] *School of Mathematics, Statistics & Actuarial Science, University of Kent* [3] *Department of Physics, Florida International University* [4] *Research Imaging Institute, University of Texas* [5] *Department of Biostatistics, University of Michigan* [6] *Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, and Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford*

November 26, 2017

Abstract

Coordinate-based meta-analyses (CBMA) allow researchers to combine the results from multiple fMRI studies 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 studies that are carried out but are not published due to lack of significance. 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 contrasts. 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 contrasts estimated as at least 6 per 100 reported.

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 (Farah, 2014). These limitations include poor test-retest reliability (Raemaekers *et al.*, 2007), excess of false positive findings (Wager *et al.*, 2009) and small sample sizes (Carp, 2012). Meta-analyses play an important role in the field of 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 xyz coordinates (foci) of peak activations that are typically published, are the main approach for the meta-analysis of 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 (Normand, 1999). 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 (Rosenthal, 1979; Iyengar and Greenhouse, 1988). This problem refers to research studies that are initiated but are not published due to lack of significance, either by cause of authors' hesitation to submit or perhaps because of rejection by journals that are reluctant to publish negative results. The file drawer along with the other forms of publication bias (see Song *et al.* (2000) for an overview) can potentially undermine the quality of a meta-analysis as they lead to biased estimates of the effect of interest (Begg and Berlin, 1988; Sutton *et al.*, 2000). Aside from distorting a particular scientific question of interest, this feeds into researchers' scepticism regarding the usefulness of meta-analysis (Greenland, 1994).

Evidence of the file drawer problem has been found in many fields of scientific research, including psychology (Kühberger *et al.*, 2014), public health (Dwan *et al.*, 2008, 2013) and the social sciences (Sterling *et al.*, 1995). Therefore, several methods have been proposed for detecting and sometimes adjusting for the presence of the file drawer problem. Early literature was focused on finding the *fail-safe N* (Rosenthal, 1979; Iyengar and Greenhouse, 1988), the minimum number of unpublished studies required to overturn the outcome of meta-analysis. Much attention has been given to the graphical tool known as the *funnel plot* (Light and Pillemar, 1984; Egger *et al.*, 1997), as well as methods that formalise this idea (Duval and Tweedie, 2000a,b, among others). Another very common approach involves the use of *weight functions* where the probability of observing a study is modelled as a function of its characteristics such as p-values, see e.g. Larose and Dey (1998); Copas and Jackson (2004). Finally, another popular approach is *sensitivity analysis* where one chooses a parametric model for the probability that an initiated body of research results in publication, and the outcome of meta-analysis is studied under different parameter values of the model (Copas, 1999, 2013). For an overview and more detailed description of methods for modelling the file drawer problem we refer the reader to Jin *et al.* (2015).

Since most of the aforementioned methodologies cannot be directly applied to fMRI coordinate-based meta-analysis data, there has been little investigation into potential biases in the field. One effort is Jennings and Van Horn (2012) that found evidence for publication biases in 74 studies of tasks involving working memory. The authors use the maximum test statistic reported in the frontal lobe as the effect estimate in their statistical tests. Another example is David *et al.* (2013), who studied the relation between sample size and the total number of activations and reached similar conclusions as Jennings and Van Horn (2012), finding publication bias mainly affecting small studies. However, to date there has been no work on estimating the fundamental file drawer quantity, that is the number of missing studies.

In what follows, we propose a model for estimating the prevalence of non-significant contrasts omitted from a large cohort of studies in the context of CBMA. 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 studies. 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

2.1 BrainMap database

Our analysis is motivated by coordinate data from *BrainMap* [1] (Laird *et al.*, 2005). BrainMap is an online, freely accessible database for coordinate-based data of both functional and structural neuroimaging experiments. The database is continuously expanding and as of November 2014 consists of results obtained from 2,562 scientific papers, each one of these containing several *experiments* or *contrasts*. BrainMap is a widely used resource, and many meta-analyses are based on data retrieved from this database (see Hill *et al.* (2014) and Kirby and Robinson (2015) for some recent examples). It is therefore of vital importance to investigate the presence of the file drawer problem and its possible effects on meta-analysis.

Our unit of observation is a contrast, and hence our dataset consists of 12,292 observations. Each observation (contrast) consists of a list of three dimensional coordinates $\mathbf{x}_i$, the *foci*, typically either local maxima or centers of mass of voxel clusters with significant activations. For the purposes of this work we ignore the spatial aspect of the problem and instead model the file drawer only based on the the total number of foci per contrast n_i. Further, we do not consider any of the resting-state studies that are registered in BrainMap. Table 1 presents presents some summary statistics of the BrainMap dataset, whereas Figure 1 shows the empirical distribution of the total number of foci per contrast.

Table 1: BrainMap database summaries.

[1] RRID:SCR_003069

Database composition

Publications	Contrasts	Foci
2,555	11,432	92,407

Contrasts per publication

Min.	Q_1	Median	Mean	Q_3	Max.
1	2	4	4.5	6	42

Foci per contrast

Min.	Q_1	Median	Mean	Q_3	Max.
1	3	6	8	11	98

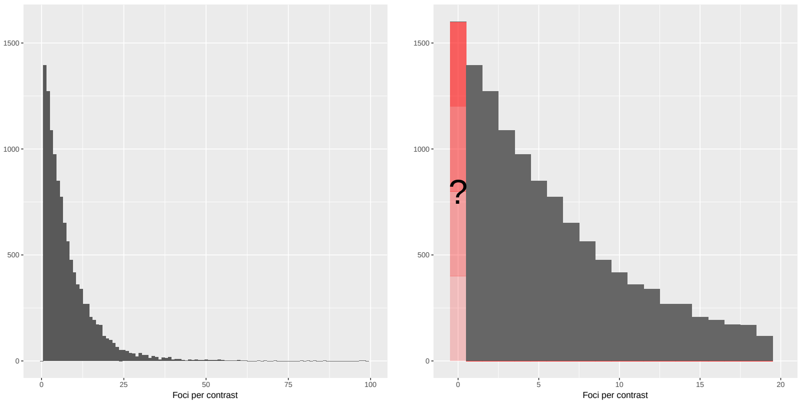

Figure 1: Empirical distribution of the total number of foci per contrast in the BrainMap database, n_i. The left panel shows the full distribution, while the right panel shows a zoomed-in view of all studies reporting 24 or fewer foci. The BrainMap database does not record incidents of null contrasts (contrasts in a study for which $n_i = 0$).

The barplot of Figure 1 (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. Thus we are careful *not* to describe our estimates as 'null study' prevalence. Rather, we are estimating prevalence of null contrasts, some of which may in fact be clearly reported in papers in the BrainMap database; however, given the stigma of negative findings, we suspect these type C null contrasts are rare.

2.2 Models

Our model uses count data from the observed, reported experiments to infer on the file drawer quantity. At this point, we make some critical assumptions: I) data $\{n_i\}_{i=1}^{I}$, 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-study covariates); II) the probability of publication depends on the total number of significant activations n_i; specifically, this probability equals zero for experiments with $n_i = 0$ and equals one for studies with $n_i \geq 1$. 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) $\pi\left(n \mid \boldsymbol{\theta}\right)$ defined on $n = 0, 1, \ldots,$ where $\boldsymbol{\theta}$ is the parameter vector, the zero truncated pmf is:

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

We consider three types of count distributions $\pi\left(n \mid \boldsymbol{\theta}\right)$: 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 set of observed points, the distribution of counts will show this over-dispersion. In particular, we note that in our previous work (Kang *et al.*, 2011) we have always used such 'Cox processes' with random intensity functions.

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:

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

where μ is the mean, $\phi > 0$ is the dispersion parameter and $\Gamma\left(\cdot\right)$ 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 $\text{Pois}(\mu\gamma_i)$ random variables; the γ_i follows a particular shifted

Gamma distribution with parameters σ and ν, $\sigma > 0$ and $0 \leq \nu < 1$ (Rigby *et al.*, 2008). The probability mass function of the Delaporte distribution can be written as:

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

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 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 homogenous data, and we can reasonably expect that differences in experiment type, sample size, etc., can introduce systematic differences between studies. 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:

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

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 2014 with median 2004; ii) the square root of the number of participants ranging from 1 to 395 with median 12; iii) the experimental context. For experimental context we use 7 levels: age effects, disease effects, drug effects, gender effects, learning, linguistic effects, normal mapping and other. The first six levels appeared at least 20 times in the database while 'other' covers any other label that occured less frequently. Even though the BrainMap database records multiple labels for each experiment, we only used the first one. 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 Rigby and Stasinopoulos

(2005). The fitting is done in R[2] (R Core Team, 2015) with the *gamlss* library (Stasinopoulos and Rigby, 2007). 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) 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 $I^*/(1 - \pi(0|\boldsymbol{\theta}))$ samples from the untruncated distributions, where I^* is chosen such that the expected number of non-zero counts is I. 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 studies is set to $I = 200$, 500, 1,000 and 2,000. For each combination of (I, μ, ϕ) and (I, 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

[2] RRID:SCR_001905

least squares in FSL[3], 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).

3 Results

3.1 Simulation results

The percent relative bias of the estimates of p_{z}, $\frac{\hat{p}_{\mathrm{z}} - p_{\mathrm{z}}}{p_{\mathrm{z}}} \times 100$, 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 $\hat{p}_{\mathrm{z}}$ is small, never exceeding 8% when the sample size is comparable to the sample size of the BrainMap database ($I = 2,555$) 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 studies, 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 $\hat{\mathbf{p}}_{\mathrm{z}}$**			
$\boldsymbol{\mu}$	$\boldsymbol{\phi}$	$\mathbf{p}_{\mathrm{z}}$	$\mathbb{E}[\mathbf{I}] = \mathbf{200}$	**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

[3] RRID:SCR_002823

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 $\hat{\mathbf{p}}_{\mathrm{z}}$**			
μ	σ	ν	$\mathrm{p_z}$	$\mathbb{E}[\mathbf{I}] = \mathbf{200}$	**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 $\hat{p}_{\mathrm{z}}$, missing experiment rate as a percentage of observed studies, 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($\hat{\mathbf{p}}_{\mathrm{z}}$)**			
μ	ϕ	$\mathrm{p_z}$	$\mathbb{E}[\mathbf{I}] = \mathbf{200}$	**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($\hat{\mathbf{p}}_z$)**			
$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\nu}$	$\mathbf{p}_z$	$\mathbb{E}[\mathbf{I}] = \mathbf{200}$	**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 2 and Table 4. In Figure 2 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, and by AIC for all 6 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.

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 modeling 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 and AIC score are shown (smaller AIC is better).

Voxelwise

		Negative Binomial		**Delaporte**	
Split	$\mathbf{n_0}$	$\mathbf{\hat{n}_0}$	**AIC**	$\mathbf{\hat{n}_0}$	**AIC**
1	7	20 [14,27]	6576.4	1 [1,4]	6562.8
2	5	22 [15,29]	6583.3	3 [1,8]	6576.2
3	5	22 [16,30]	6575.9	1 [1,21]	6566.1
4	4	25 [18,33]	6603.8	4 [1,25]	6601.7
5	10	18 [13,24]	6539.4	1 [0,1]	6504.1
6	10	21 [15,29]	6557.0	2 [1,10]	6550.0

Clusterwise

		Negative Binomial		**Delaporte**	
Split	$\mathbf{n_0}$	$\mathbf{\hat{n}_0}$	**AIC**	$\mathbf{\hat{n}_0}$	**AIC**
1	148	167 [115,248]	3167.8	71 [50,108]	3166.9
2	144	151 [104,217]	3209.3	58 [44,79]	3204.5
3	150	161 [109,246]	3163.2	95 [62,184]	3164.4
4	151	154 [107,231]	3156.3	106 [57,159]	3158.0
5	153	148 [101,291]	3175.0	98 [60,174]	3176.5
6	152	151 [103,223]	3198.2	89 [54,157]	3199.7

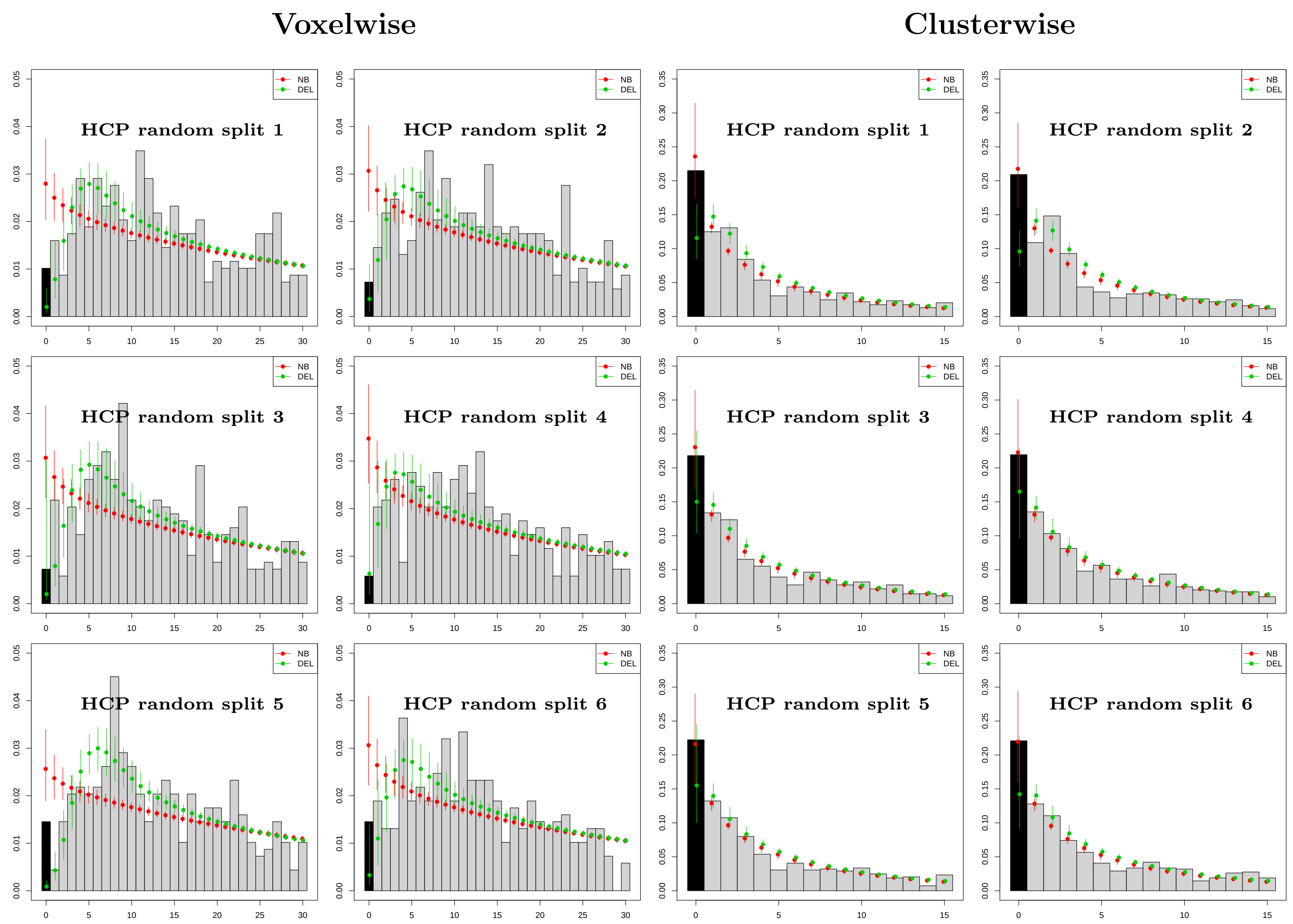

Figure 2: Evaluation with HCP data with 688 contrasts of sample size 10, comparing accuracy of Negative Binomial (NB) and Delaporte (DEL) distributions for the prediction of the number of studies with no significant results (zero foci) based on only significant results (one or more foci). Left panel 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 6), and we do not consider it further. We start by fitting the Negative Binomial and Delaporte zero-truncated models without any covariates. Figure 3 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 but with little difference in AIC (Table 6). 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 studies 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	$\hat{\mathrm{p}}_{\mathrm{z}}$	**95% interval**	$\hat{\mathrm{p}}_{\mathrm{z}}$	**95% interval**
A	10.14	[8.55,12.00]	7.27	[4.74,11.03]
B	9.47	[8.00,11.11]	5.41	[3.92,8.95]
C	9.02	[7.61,10.63]	6.17	[4.10,9.77]
D	8.67	[7.31,10.21]	7.03	[4.41,9.67]
E	10.16	[8.59,11.97]	6.76	[4.67,10.60]

Covariates essentially have no effect on the estimated prevalence of missing contrasts. However, including covariates results in an improvement in terms of AIC for both models and all BrainMap subsamples (Table 6). As can be seen in Figure 4, 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. Finally, we 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 5).

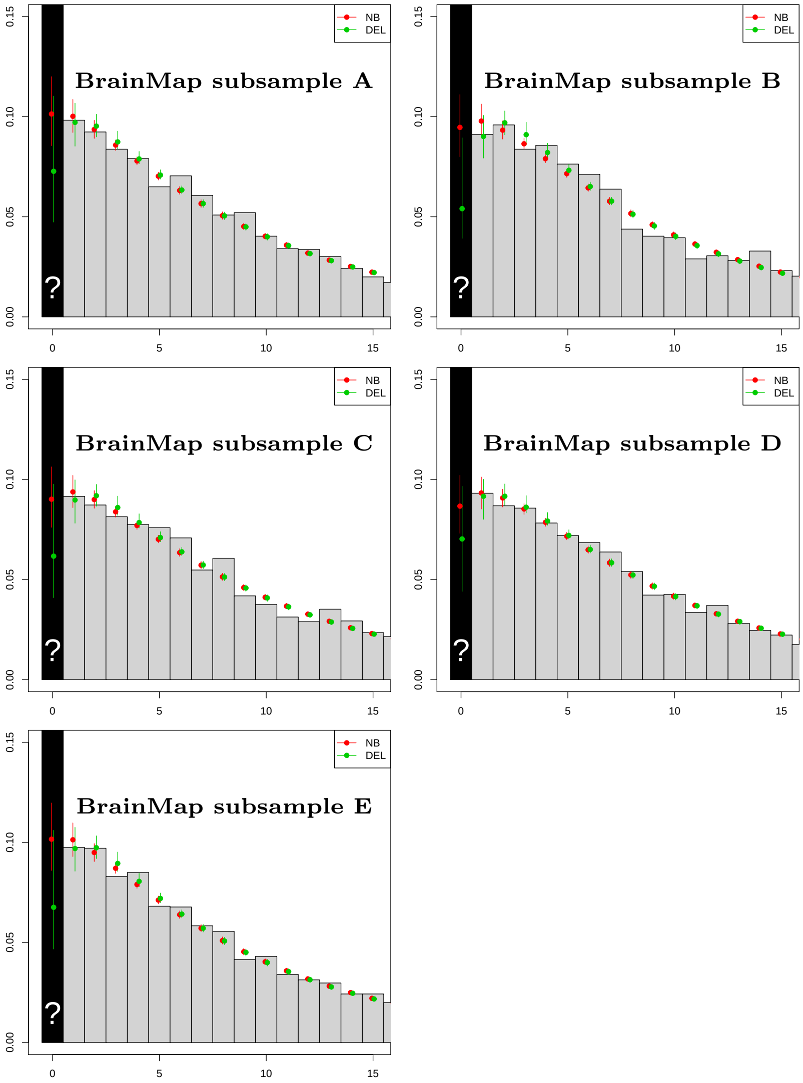

Figure 3: BrainMap results for 5 random samples using the Negative Binomial and Delaporte models and no covariates. Plots show observed count data (gray bars) with fit of full (non-truncated) distribution based on zero-truncated data, including the estimate of p_0 (over black bar).

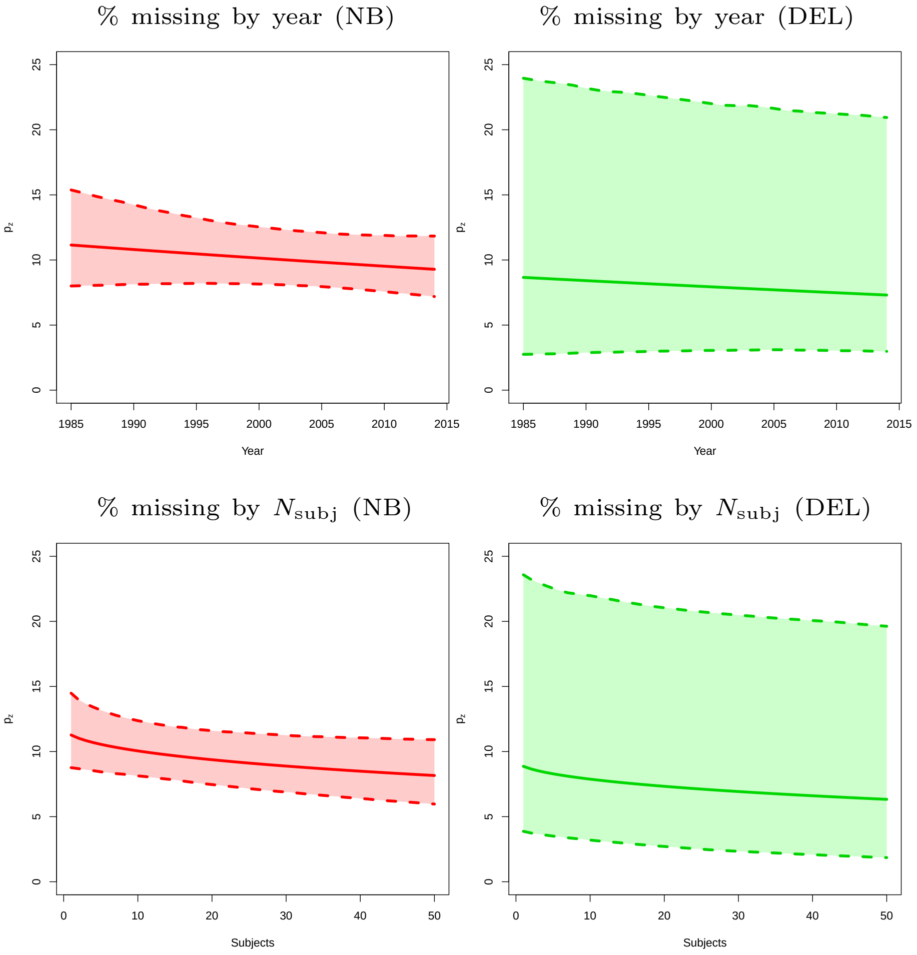

Figure 4: Predicted p_z, missing experiment rate per 100 published experiments, as a function of year of publication (top) and the square root of sample size (bottom), with pointwise 95% bootstrap confidence intervals. There is not much variation in the estimate of the percentage missing, but in both 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 first BrainMap random sample (subsample A).

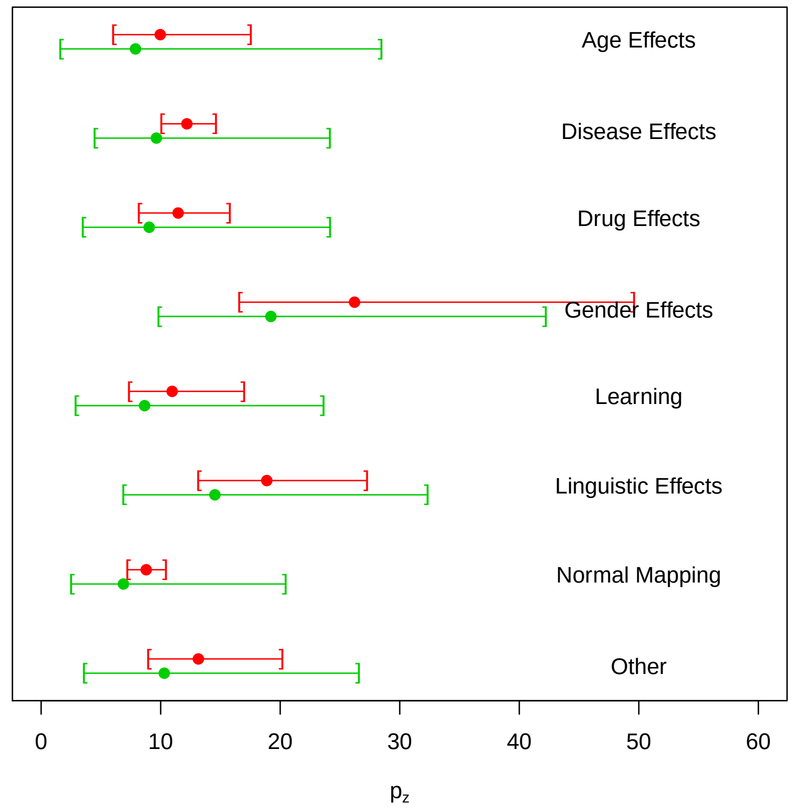

Figure 5: Studies missing per 100 published as a function of experiment context, with 95% bootstrap confidence intervals. Note that we have fixed the year and square root sample size covariates to their median values. The plot is derived from the first BrainMap random sample (subsample A).

Table 6: AIC model comparison results for the BrainMap data. The AIC is found by fitting 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 covariates (smaller AIC indicates better fitting model).

AIC model comparison: Negative Binomial

Subsmaple	**No covariates**	**Regression**
A	16111.28	15095.05
B	16012.85	14994.86
C	16206.25	15178.73
D	16056.65	15037.79
E	16005.15	15006.94

AIC model comparison: Delaporte

Subsmaple	**No covariates**	**Regression**
A	16112.77	15082.81
B	16011.35	15014.12
C	16207.33	15185.88
D	16058.47	15028.52
E	16006.02	15036.04

4 Discussion

Summary. In this paper, we have proposed a method for estimating the total number of contrasts missing from a large cohort of CBMA studies due to the file drawer problem. Our method uses intrinsic statistical characteristics of the non-zero count data to infer zero counts. This is achieved by estimating the parameters of a zero-truncated model, either Negative Binomial or Delaporte, which are susequently used to predict the prevalence p_0 of zero-count studies in the original, untruncated distribution, and re-expressing this as p_z, the rate of missing contrasts per 100 published. The approach 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 experiments with at least one significant activation are always published. The assumption that non-significant research findings are suppressed from the literature has been adopted by authors in classical meta-analysis (Eberly and Casella, 1999, among others) and we believe that is reasonable in the context of CBMA as well.

A series of simulations studies suggest that the zero-truncated modelling approach provides valid estimates of the file drawer quantity. 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 studies decreases and therefore the methods should not employed for meta-analyses with a small number of studies, below, say, 1,000 studies.

We find that both zero-truncated Negative Binomial and Delaporte models provide a good fit for the total number of foci per contrast in the BrainMap database. The analysis suggests that the estimated magnitude of the file drawer slightly varies depending on study characteristics, but generally consists of at least 6 missing experiments for 100 published and is significantly greater than zero.

Implications for meta-analyses. Our findings provide evidence for the existence of publication bias in CBMA. While we cannot rule out a contribution of contrasts that have actually been reported in the original publications and not encoded in the database, we posit that the majority come from contrasts never described in publications or not published at all.

We stress that this analysis is totally agnostic to the statistical procedures used to generate the results in the BrainMap 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 studies. In the end, our results reflect the aggregate statistical practice and signal structure reflected in the BrainMap database.

Ideally, our unit of inference would be a publication. 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 of non-null contrasts.

The presence of missing experiments does not invalidate existing studies, but complements the picture seen when conducting a literature review. Nevertheless, there are some implications concerning the interpretation of the results obtained from current CBMA approaches. In particular, methods that make no adjustments for the file drawer effect are conditional on the existence of at least one activation. Hence, effect estimates obtained trough a CBMA are inflated unless the prevalence of null exeriments is zero.

Future work. 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[4] (Yarkoni *et al.*, 2011), though note Neurosynth does not report

[4] RRID:SCR_006798

foci per contrast but per paper). Secondly, one may argue that our censoring mechanism is rather simplistic, and does not reflect the complexity of current (and potentially) poor scientific practice. For example, 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 of initially-censored (zero-count) data being 'promoted' to a non-zero count through some means. Such models can be fit under the Bayesian paradigm and will be examined in our future work.

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 work presented in this paper represents the views of the authors and not necessarily those of the NIH or the Wellcome Trust Foundation.

References

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

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

Copas, J. (1999). What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **162**(1), 95–109.

Copas, J. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **62**(1), 47–66.

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

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

Duval, S. and Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, **95**(449), 89–98.

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

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., and Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*, **3**(8), e3081.

Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., and the Reporting Bias Group (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias - An updated review. *PLoS ONE*, **8**(7), e66844.

Eberly, L. E. and Casella, G. (1999). Bayesian estimation of the number of unseen studies in a meta-analysis. *Official Journal of Statistics*, **15**(4), 477–494.

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

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

Greenland, S. (1994). Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, **140**(3), 290–296.

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

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

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

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

Kang, J., Johnson, T. D., Nichols, T. E., and Wager, T. D. (2011). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *Journal of the American Statistical Association*, **106**(493), 124–134.

Kirby, L. A. and Robinson, J. L. (2015). Affective mapping: An activation likelihood estimation (ALE) meta-analysis. *Brain and Cognition*. In press.

Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, **9**(9), 1–8.

Laird, A. R., Lancaster, J. J., and Fox, P. T. (2005). Brainmap: the social evolution of a human brain mapping database. *Neuroinformatics*, **3**(1), 65–77.

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

Light, R. J. and Pillemar, D. B. (1984). *Summing up: the science of reviewing research*. Harvard University Press.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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

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

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

Song, F., Eastwood, A. J., Gilbody, S., Duley, L., and Sutton, A. J. (2000). Publication and related biases. *Health Technology Assessment*, **4**(10), 1–191.

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

Sterling, T. D., Rosenbaum, W. L., and Weinkam, J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, **49**(1), 108–112.

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

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., and Van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, **45**(Supplement 1), S210–S221.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, **8**(8), 665–670.

A BrainMap summaries for study 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 study context. In particular, Table 7 presents the total number of studies per level, the average sample size per contrast, and the average number of foci per contrast. Note that in subsamples C and E there were less than 20 contrasts with label 'Gender effects'; hence, we incorporate those in the 'Other' category.

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

Contrasts per level					
Study context	**BrainMap subsample**				
	A	**B**	**C**	**D**	**E**
Age effects	28	20	30	25	25
Disease effects	472	453	463	487	468
Drug effects	56	52	55	61	60
Gender effects	21	21	-	21	-
Learning	27	28	31	31	31
Linguistic effects	30	32	28	36	32
Normal mapping	1736	1768	1746	1717	1739
Other	25	21	43	18	41
Average constrast sample size					
Study context	**BrainMap subsample**				
	A	**B**	**C**	**D**	**E**
Age effects	15.3	14.1	11.6	13.0	14.1
Disease effects	13.4	13.4	14.1	14.1	13.4
Drug effects	12.7	12.9	12.3	11.9	12.1
Gender effects	12.3	12.6	-	12.5	-
Learning	11.1	11.0	12.1	12.3	11.9
Linguistic effects	11.3	11.8	11.3	11.6	12.1
Normal mapping	13.3	13.3	13.3	13.3	13.3
Other	11.4	11.3	12.5	11.2	12.1
Average foci per contrast					
Study context	**BrainMap subsample**				
	A	**B**	**C**	**D**	**E**
Age effects	8.8	8.5	9.5	8.8	9.4
Disease effects	7.6	7.4	7.6	7.4	7.2
Drug effects	8.1	9.0	8.6	8.8	8.7
Gender effects	4.3	5.2	-	3.6	-
Learning	8.0	8.5	8.4	8.4	8.1
Linguistic effects	5.3	8.6	7.8	10.1	7.2
Normal mapping	9.8	9.5	9.9	9.7	9.6
Other	7.0	7.8	6.1	5.8	5.2

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 6. It is evident that the zero-truncated Poisson model provides a poor fit to the BrainMap data. The finding is confirmed by the AIC criterion which is 26067.2, 25303.4, 26006.6, 25018.8 and 25507.7 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 studies is estimated as almost zero in all subsamples (Figure 6, final plot). However, these estimates should not be trusted considering the poor fit provided by the zero-truncated Poisson model.

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	μ	ϕ	μ	σ	ν
A	8.28	0.89	8.50	0.93	0.046
B	8.19	0.85	8.50	0.96	0.088
C	8.52	0.84	8.75	0.90	0.054
D	8.33	0.81	8.46	0.84	0.031
E	8.12	0.88	8.38	0.94	0.060

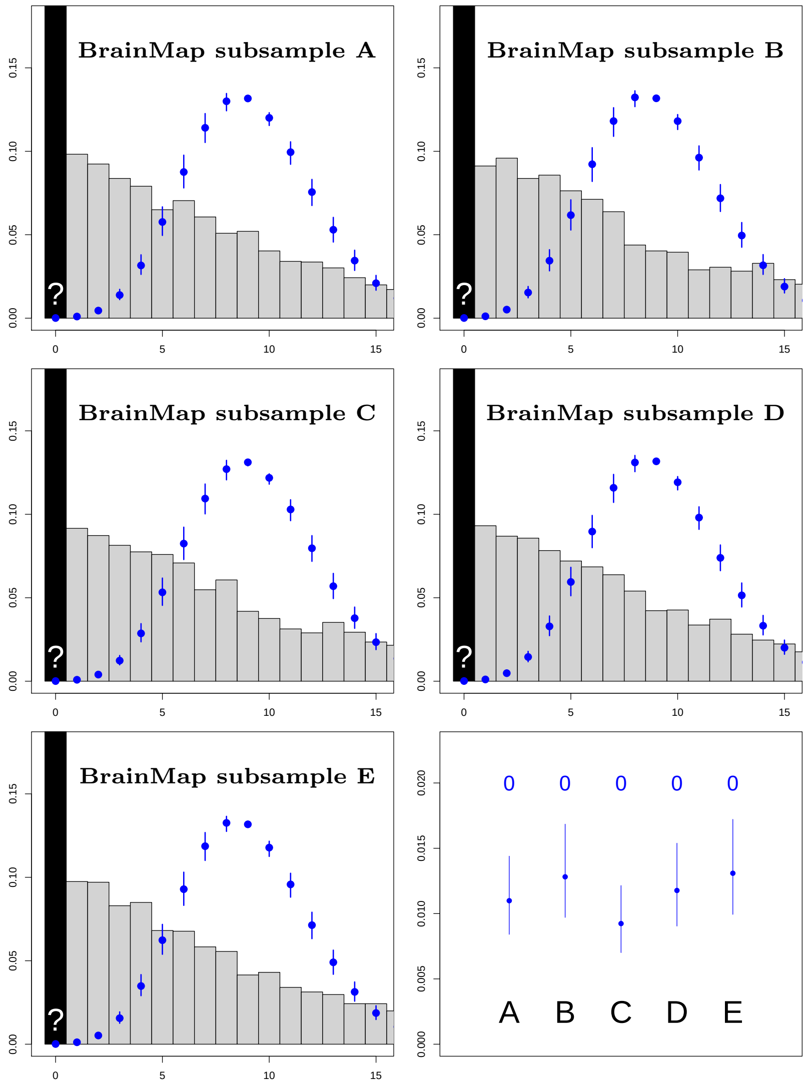

Figure 6: BrainMap results for 5 random samples using the zero-truncated Poisson distribution. The first 5 plots show observed count data (gray bars) with fit of full (non-truncated) distribution based on zero-truncated data, including the estimate of p_0 (over black bar). Final plot shows estimates of p_z, prevalence of file drawer studies for every 100 studies observed. All fitted values include 95% bootstrap confidence intervals. The Poisson model provides a poor fit to all 5 subsamples.