



UNIVERSITAT_{DE}
BARCELONA

Development, implementation, and validation of a new method for meta-analysis of voxel-based neuroimage studies

Anton Albajes-Eizagirre



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- Compartitqual 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - Compartitqual 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-ShareAlike 4.0. Spain License.**



UNIVERSITAT_{DE}
BARCELONA

**DEVELOPMENT, IMPLEMENTATION, AND
VALIDATION OF A NEW METHOD FOR
META-ANALYSIS OF VOXEL-BASED
NEUROIMAGE STUDIES.**

by

Anton Albajes-Eizagirre

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Directed by

Joaquim Radua

Edith Pomarol-Clotet

Tutored by

David Bartres Faz

PROGRAMA DE DOCTORAT EN MEDICINA I RECERCA
TRANSLACIONAL

December 2020

Declaration of Authorship

El Dr. Joaquim Radua, doctor en medicina per King's College London, University of London i la Dra. Edith Pomarol-Clotet, doctora en medicina per la Universitat de Lübeck

CERTIFIQUEN

que la memòria 'Development, implementation, and validation of a new method for meta-analysis of voxel-based neuroimage studies', presentada per Anton Albajes-Eizagirre, ha estat realitzada sota la nostra direcció i considerem que reuneix les condicions necessàries per ser defensada davant el Tribunal corresponent per optar al Grau de Doctor.

Dr. Joaquim Radua

Imaging of mood- and anxiety-related disorders

Institut d'Investigacions Biomèdiques August Pi i Sunyer

Dra. Edith Pomarol-Clotet

FIDMAG Germanes Hospitalàries Research Foundation

“Gran és aquell qui mai no ha perdut el seu cor de nen.”

Mencius

Acknowledgements

A l'atzar agraeixo haver comptat amb l'ajuda de tanta gent que ha tingut la generositat i la paciència d'espentejar-me en el camí de la ciència i fins aquí. Al Josep Maria Miret, que em va descobrir la passió per la recerca. A l'Aureli Soria-Frisch, que em va acollir a tants projectes seus. A l'Edith Pomarol-Clotet, per l'ineestimable i imprescindible suport durant gran part d'aquest projecte. I, per descomptat, al Joaquim Radua, autèntica ànima i cervell d'aquesta tesi que tan generosament ha volgut compartir amb mi.

A la meva germana i els meus pares, per haver-me donat absolutament tot el que he necessitat en tots els aspectes, i moltíssim més, en aquesta vida privilegiada que gaudeixo.

A la Maria, que fa possible que malgrat tots els meus defectes, me n'acabi més o menys sortint.

This work was supported by PFIS Predoctoral Contract FI16/00311.

Contents

Declaration of Authorship	i
Acknowledgements	iii
List of Figures	v
1 Introduction	1
1.1 Neuroimage, a young and prosper area	1
1.2 The replication crisis	1
1.3 Publication and other reporting biases	5
1.4 How the replication crisis, publication biases, and other issues affect neuroimage studies	10
1.5 Meta-analyses are very powerful to confront these challenges	12
1.6 Meta-analysis in voxel-based neuroimaging	14
1.7 Historic of methods and their problems	15
1.7.1 Kernel-Based	16
1.7.2 Model-based	17
1.8 AES-SDM as our starting point	18
2 Objectives	19
3 Articles included in this thesis	21
4 Discussion	59
5 Conclusions	63
References	64

List of Figures

1.1	Number of studies referenced in PUBMED between the years 2000 and 2019 searching by keywords 'Fractional anisotropy' or 'Voxel' or 'fMRI' or 'Neuroimaging' or 'Gray matter volume' or 'Bold response'	2
1.2	Three examples of the checking replication comparing thresholded maps. In the first case, while study A1 and study A2 have very similar results, being A1 slightly positive and A2 slightly negative the replication is considered failed. In the second case, while study B1 is almost significant, study B2 is non-statistically significant and their results are very different, the replication is considered successful. In the third case, although C1 is considered significant, the shape of the curve makes the results suspicious of being spurious. Nonetheless, the replication while compared with the solid result of the study C2 is considered successful.	5

Chapter 1

Introduction

1.1 Neuroimage, a young and prosper area

Neuroimaging as the research area that we know today was born in the early 1970 s with the invention of computerized tomography (CT) and the early 1980s with the invention of the positron emission tomography (PET) and the magnetic resonance imaging (MRI). It is, therefore, a relatively young but flourishing research area ([Leeds and Kieffer \(2000\)](#), [Filler \(2009\)](#)) in which there was a quick growth until a few years ago (see [Figure 1.1](#)). On the bright side, this increasing activity talks about the prominence that neuroimaging research is attaining. On the other side, the consolidation of neuroimaging as a ‘rightful’ research area comes with the obstacles and challenges of any discipline that aims to achieve high-quality scientific standards. Among all the obstacles present in the process of consolidation of new knowledge in modern sciences, I will firstly discuss two that sensibly affect neuroimaging research and for which the usage of the meta-analysis techniques presented later in this work can be more beneficial: the replication crisis and the existence of publication bias.

1.2 The replication crisis

It was Karl Popper who stated that “non-reproducible single occurrences are of no significance to science” ([Popper \(1934\)](#)), reminding us that reproducibility is a fundamental pillar of the scientific method ([Repko and Szostak \(2020\)](#)). Popper also raised the questions: What is reproducibility? And replicability? According to the American Statistical Association (ASA), “A study is reproducible if you can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study” ([ASA Develops Reproducible Research Recommendations \(n.d.\)](#)). With

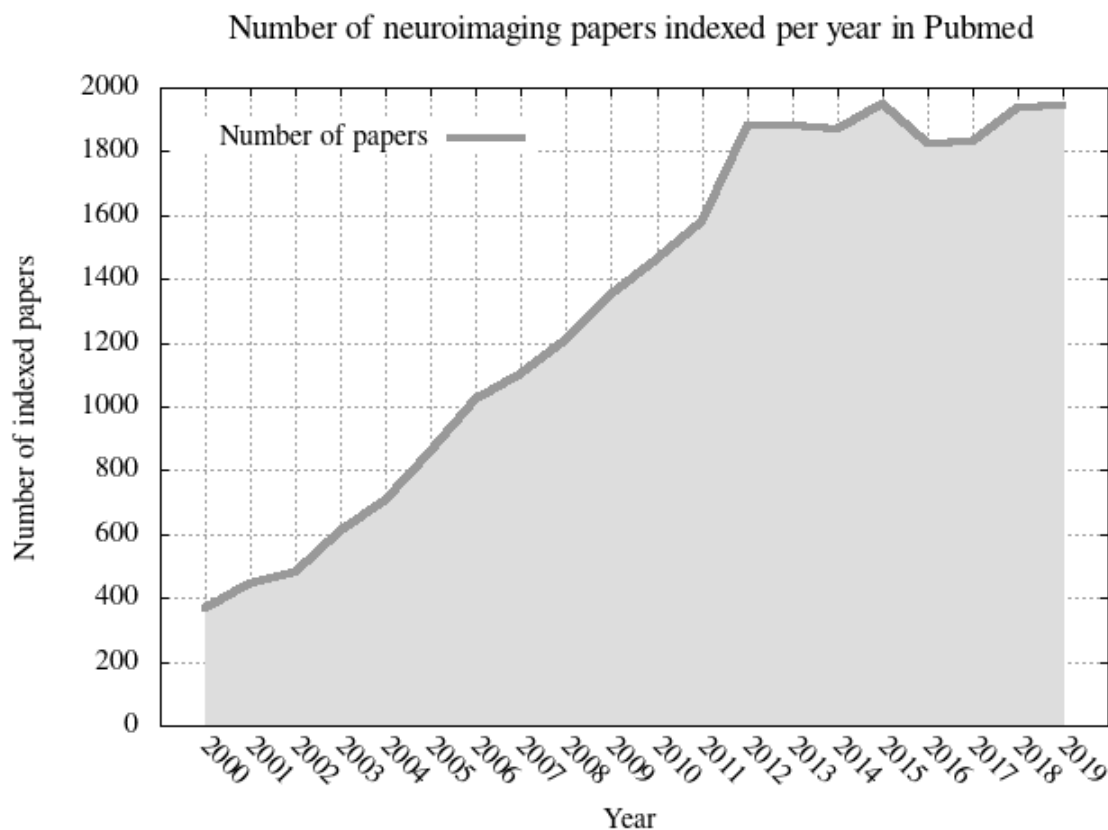


FIGURE 1.1: Number of studies referenced in PUBMED between the years 2000 and 2019 searching by keywords 'Fractional anisotropy' or 'Voxel' or 'fMRI' or 'Neuroimaging' or 'Gray matter volume' or 'Bold response'

regards to replicability, ASA defines “the act of repeating an entire study, independently of the original investigator without the use of original data (but generally using the same methods)” as replicability. Furthermore, what constitutes a corroboration? According to Karl Popper, corroboration is “a qualitative measure of the success that a theory has achieved in passing tests; that is, that the predictions made from it are borne out by experience or experiment, at least as well can be determined under the conditions and precision of the tests and experimenters” (*Improving Reproducibility and Replicability - Reproducibility and Replicability in Science - NCBI Bookshelf* (n.d.)). I think it is safe to conclude that, while corroboration constitutes the ultimate confirmation of a scientific finding or knowledge, it shall be achieved by repeatedly reproducing novel results (if possible due to the availability of original conditions) and later replicating such results in different or new conditions. Moreover, while reproducing does not necessarily ensure the validity of the original results, it is useful to improve research quality by verifying findings and identifying weaknesses, errors or even questionable research practices (Harris et al. (2019)).

But replication is scarcely practiced by the scientific community and this lacuna has

become a growing concern. In a poll conducted by the journal *Nature* ([Baker \(2016\)](#)) in 2016, 70% of asked researchers recognized that they failed to replicate at least one published work, and even 50% of them failed to reproduce a published work of their own. Moreover, in the same poll, 90% of researchers agreed that the replication crisis is real. However, only a minority reported to have tried to publish a replication study, and only less than 20% reported to have been contacted by other researchers regarding failure to replicate their work. Causes for low replicability are various and can be differentiated regarding what they affect the most. According to the researchers polled by *Nature*, causes for the low replicability of modern scientific work are selective reporting, low statistical power, lack of publication options for methods or original data and, less frequently, fraud. According to a survey of public health analysts ([Harris et al. \(2018\)](#)), only 14% of surveyed public health researchers had shared their data and/or code. In light of these figures, sharing original data and methodologies have been proposed as a strong improvement of reproducibility and replicability of health sciences studies.

Other factors are also argued as causes for low reproducibility ([Szucs and Ioannidis \(2017\)](#)), like an estimated high ratio of false-positive results, which is a factor also related to the publication bias, the other issue that I will address in detail later. To confront this spurious positive results matter, dropping the significance threshold for discoveries to 0.005, instead of the current commonly used value of 0.05, has been proposed ([J. P. A. Ioannidis \(2005\)](#), [Benjamin et al. \(2018\)](#)). This would enable the discarding of dubious positive results and, expectedly, select only the most robust ones. However, as I will comment later, this would only apply to those attempts in which to consider if a replication is successful only the coincidence of results while applying the same significance threshold is considered. Setting our attention back to the replication crisis rather than replicability of particular studies, one of the main causes of the low levels of studies replication may be related to how the scientific world is constructed around an incentive scheme in a hypercompetitive academic climate. In what is often called the publish or perish system, the career sustainability of a researcher depends on his or her ability to maintain a constant flow of publications. Moreover, this Damoclean criteria piles on top of usually very precarious economic conditions for researchers that tend to make them take more prudent approaches and choose to conduct research work more likely to be published. At the end of the day the system rewards the quantitative amount of published work, sometimes more than the quality or the methodological neatness ([Aitkenhead \(2013\)](#)). Furthermore, some authors have analyzed the relation of this publish or perish system to the gender bias in engineering sciences. They found that while women publish their works in journals with higher impact, men dominate 80% of the publication volume, receive more citations and get more funding ([Ghiasi, Larivière, and Sugimoto \(2015\)](#)).

This skewed-scheme, together with the scant reward given to non-original works (i.e. replications) and the considerably lower chance of getting replication works published, makes researchers prefer, even if not consciously (John, Loewenstein, and Prelec (2012)), focusing on novel findings rather than on trying to replicate third parties results that may not yet be consolidated. Some studies (Niven et al. (2018), Aarts et al. (2015), Begley and Ellis (2012)) have tried to estimate the reproducibility of published works in specific areas, and it attained sensibly different rates among areas. Nevertheless, some of them were concerningly low: around 40% in psychology, 10% in cancer biology, or 40% in clinical research, for example.

This concern is not only shared among the scientific community (Munafò et al. (2014)) but has also emerged to the general media. Major news outlets, like The Guardian (Kirchherr (2017)) or the New York Times (Johnson (2014)), have featured articles about the replication problem and how it affects the credibility of science and its reliability in the eyes of the society. After all, reproducibility and science credibility have been closely linked since as early as the 17th century (Shapin (1989)).

However, the concept of replicability is also disputed. Or, rather, the concept of replication, what constitutes a replication, especially for neuroimaging studies. As stated earlier, replication refers to the ability to obtain the same results as in a reference work. But often neuroimaging results are thresholded using a discretionary threshold of $p < 0.05$ corrected for multiple comparisons. Replication implies, therefore, comparing thresholded maps in which every region not reaching this threshold is considered as a negative result. No matter how close it was to become a positive result. As seen in figure 1.2, plots A1 and A2, in practice this may very well mean that nearly positive results in one region of one study are not considered to match barely positive results in the same region of another study, and vice versa. This results in many false mismatches in regions with values close to the chosen threshold when, in fact, the real differences may be very tiny. In summary, the common practice of thresholding the results in neuroimaging studies is likely to have a high impact on the replicability and even the reproducibility of such studies. And a recent study showed that it has, indeed (Botvinik-Nezer et al. (2020)). Therefore, the problem may be, to some degree, more related to the tools and techniques employed to conduct the replicability analysis than to the original results.

The scientific community has been responding to the concern for the replication crisis and developing strategies to counteract the trend (Harris et al. (2019)). Journals like Science and others have opened new sections dedicated to replication studies (*Science/AAAS — Special Section: Data Replication and Reproducibility* (n.d.), Pashler and Wagenmakers (2012)) or published dedicated special issues like the virtual special issue on replication by the Elsevier editorial (*Virtual Special Issue On Replication Studies* (n.d.)). Since 2017

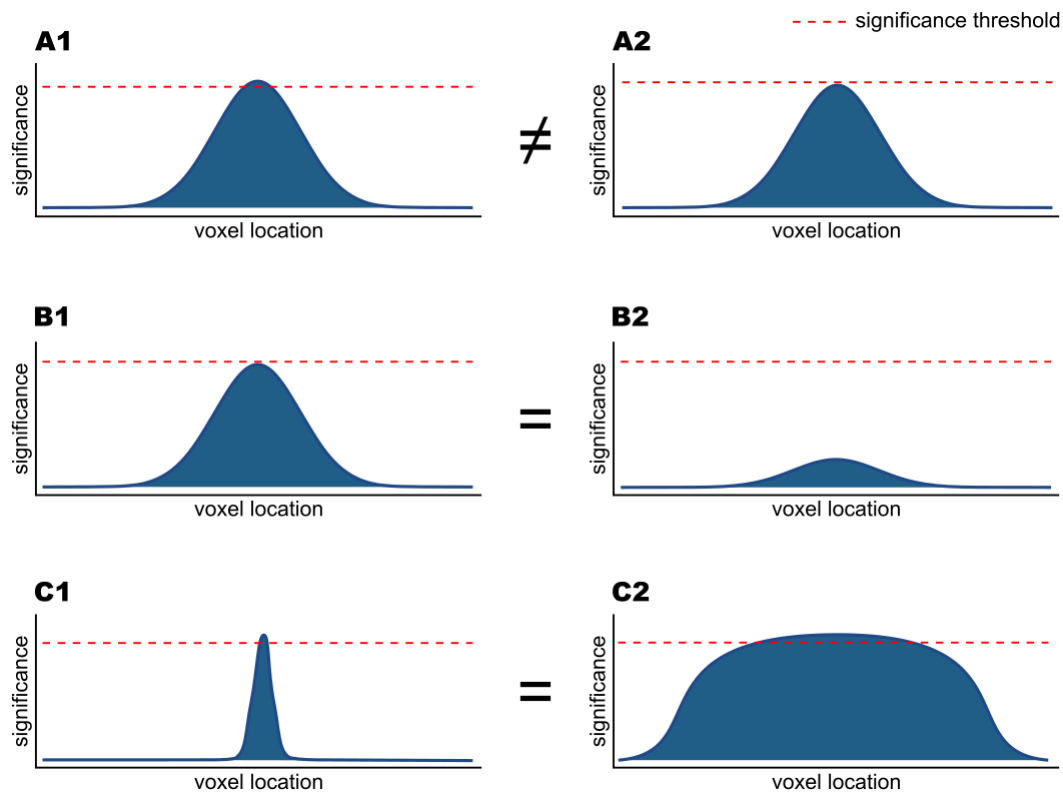


FIGURE 1.2: Three examples of the checking replication comparing thresholded maps. In the first case, while study A1 and study A2 have very similar results, being A1 slightly positive and A2 slightly negative the replication is considered failed. In the second case, while study B1 is almost significant, study B2 is non-statistically significant and their results are very different, the replication is considered successful. In the third case, although C1 is considered significant, the shape of the curve makes the results suspicious of being spurious. Nonetheless, the replication while compared with the solid result of the study C2 is considered successful.

the Organization for Human Brain Mapping (OHBM) annually presents the award to the best replication study in neuroimage (*Replication Award - www.humanbrainmapping.org* (n.d.)), and in 2017, the IRIS digital repository established a bi-yearly award to the best publication of a replication study using material held on their repository (*IRIS Replication Award- IRIS Digital Repository* (n.d.)).

1.3 Publication and other reporting biases

Another common cause of concern in modern science is the issue of publication bias and related reporting biases. Publication bias occurs when the chances of a research study being published depend on factors other than scientific quality. A common factor causing publication bias is the direction of the detected evidence: Analogously to the aforementioned problem of the scant reward granted to replication studies, the reward for negative results is also very dim. Studies reporting negative results are less attractive

to journals and are harder to get published, causing what is called the “file drawer effect” (Rosenthal (1979)). An informative systematic review (Hopewell, Loudon, Clarke, Oxman, and Dickersin (2009)) found five studies that investigated the association between the direction of their results with publication chances. Trials with positive findings were more likely to be published than trials with negative or null findings, with an odds ratio of 3.90 (95% CI: 2.68,5.68), corresponding to a risk ratio of 1.78 (95% CI: 1.58, 1.95). Another study (Song et al. (2009)) performed a systematic review of studies that tracked a cohort of studies and reported the odds of formal publication by study results, with the addition of analyzing the effect of the starting time of follow-up in the trials. They analyzed the presence of publication bias by calculating the odds ratio of publication of studies with positive results, compared to those without positive results in four subgroups: in cohorts that followed the research from inception the ratio was of 2.78 (95% CI: 2.10 to 3.69); in trials submitted to regulatory authority was of 5.00 (95% CI: 2.01 to 12.45); in abstract cohorts was of 1.70 (95% CI: 1.44 to 2.02); and in cohorts of manuscripts was of 1.06 (95% CI: 0.80 to 1.39). Altogether, the results showed a clear preference for trials with positive results as compared to trials with negative results.

One of the studies that suggested lowering the significance threshold to improve replication indices (J. P. A. Ioannidis (2005)), analyzed the specific factors making a study with positive results more likely to be published (*Publication bias - Wikipedia* (n.d.)):

- The studies conducted in a field have small sample sizes.
- The effect sizes in a field tend to be smaller.
- There is both a greater number and lesser preselection of tested relationships.
- There is greater flexibility in designs, definitions, outcomes, and analytical modes.
- There are prejudices (external interests).
- The scientific field is hot and there are more scientific teams pursuing publication.

Along with publication bias, there are other related reporting biases, like the reporting bias called “time-lab bias”. Even in the comparatively infrequent cases that get published, studies with negative results often take more time to get published than studies with positive results, causing the time-lag bias (Ioannidis (1998)). A study analyzed the time to the publication of 196 clinical trials and the influence of the direction of the results. The authors found that trials with positive findings tended to be published after 4-5 years, while it took 6-8 years for trials with negative findings to be published (Hopewell, Clarke, Stewart, and Tierney (2007)). Although this reporting bias has been identified on numerous occasions (Decullier, Lhéritier, and Chapuis (2005), Suñé, Suñé,

and Montoro (2013), Qunaj et al. (2018)) transversal to all science areas, a study found no evidence of it while analyzing the effect of the direction of the results on the time elapsed between trial completion and publication in 208 clinical trials published in BMJ, Lancet, New England Journal of Medicine and JAMA (Jefferson et al. (2016)). Nevertheless, the literature supports that time-lag reporting bias may be more present in other journals and should be considered while reviewing all published evidence to consolidate new scientific knowledge (J. P. Ioannidis, Munafò, Fusar-Poli, Nosek, and David (2014)).

There is another reporting bias that occurs when the publication of a new finding attracts the quick publication of a study with contradictory findings, and is called the “Proteus phenomenon”. Perhaps provoked by tantalizing editors (J. P. Ioannidis et al. (2014)), this results in extreme results published rapidly right after the first publication of the new finding (Pfeiffer, Bertram, and Ioannidis (2011), J. P. Ioannidis and Trikalinos (2005)). This oscillation creates a bias associating the strength of the finding with the time of publication; even if the directions of the results are contradictory, the strengths of them may be higher in earlier studies in the presence of the Proteus bias.

And yet another reporting bias appears on occasions when studies analyze different outcomes but only report a selection of them, or report incompletely the results of some of the outcomes. This elective reporting causes what is called the “outcome reporting bias”. From the analysis of 283 Cochrane reviews of clinical trials (Kirkham et al. (2010)), it was found that 55% of the reviews did not include all data for the review primary outcome. Among the included trials in the reviews, 6% of them were identified as having not reported any or some of the results of the primary outcome. Among the reviews, 34% of them included at least one trial with high suspicion of outcome reporting bias for the primary outcome.

The publication bias and these other reporting biases presented here may have an important impact on the process of consolidating new scientific evidence (J. P. Ioannidis et al. (2014)). The results of reviews on published results can be severely affected by the presence of such biases on the included publications (Guyatt, Oxman, Montori, et al. (2011), Szucs and Ioannidis (2017)). To detect the potential presence of such biases and to correct, if possible, their effect, some actions can be taken, and some techniques can be applied. Except for selective outcomes, which sometimes can be identified and corrected at the study level, potential publication and reporting bias can only be detected at the group level. That is, analyzing a group of studies and detecting the potential presence of bias. A common technique to detect the presence of potential publication bias is to assess if there is asymmetry on the scatter plot displaying the reported effect size against some measure of study precision (J. Light and B. Pillemer (1986), Sterne and Egger (2001)). These plots are called funnel plots, and can be visually inspected (Al et

(2016)) or statistically tested, but this technique requires a cautious application, as the interpretation of funnel plots can be prone to error in some cases (Terrin, Schmid, and Lau (2005), Lau, Ioannidis, Terrin, Schmid, and Olkin (2006)) (like when there is high between-study heterogeneity). A study analyzed the reliability of funnel plots for publication bias detection through asymmetry assessment (Kicinski (2014)). Even though they simulated sets of studies with up to four times more positive results than other outcomes, the test based on the funnel plot asymmetry detected the publication bias on only 15% of the cases. An important aspect of such simulations is that in addition to a high number of positive significant results, they also simulated a high between-study heterogeneity. Although other quantitative methods relying on the same principle of the relationship between effect size and study precision have been developed (Begg and Berlin (1988), Egger, Smith, Schneider, and Minder (1997)) and are becoming commonly used, there are still some cases in which they need to be taken with precaution (Irwig, Macaskill, Berry, and Glasziou (1998), Stuck, Rubenstein, and Wieland (1998)). More methods are being developed to overcome these issues (Mueller et al. (2013), Lin and Chu (2018)).

Concerning the time-lag bias, detecting its potential presence in a set of publications may be complex, especially if most of the included studies share approximately the same publication times. Young research areas or novel findings are likely to produce scientific works closely grouped in time. In such cases, there are not many strategies to be taken beyond taking the results prudently and waiting for new studies to be published. However, if the publication dates of the studies span over a relatively large time, a regression can be applied to check the effect of the publication date (Tsujimoto et al. (2017)). Another proposed approach is to develop a cumulative group analysis in which an initial subset of earlier studies is included to perform the analysis, and then repeat the analysis several times incorporating each time newer studies into the set (Lau et al. (1992)). The idea is that if the effects of the analysis show a trend, it may be an indication of time-lag bias.

Lastly, as for the outcome reporting bias I already stated earlier that sometimes this bias can be detected at the study level (Guyatt, Oxman, Vist, et al. (2011)). However, the possible presence of outcome reporting bias in the results of the studies included in a collection can be analyzed with what is called an excessive significance test (J. P. Ioannidis and Trikalinos (2007)). The principle behind the test, which may also detect other biases, is to calculate the expected number of studies with statistically significant results according to their statistical power. If this expected number of statistically significant studies is sensibly smaller than the observed number of statistically significant studies in our set, this may be an indication of the presence of outcome reporting bias. For this

principle to sustain, effect sizes need to be relatively homogeneous. Therefore, the reliability of the test results needs to be assessed in each application, and it might depend on the conditions of each set of studies: “In the presence of considerable between-study heterogeneity, efforts should be made first to dissect sources of heterogeneity. Applying the test ignoring genuine heterogeneity is ill-advised” (J. P. Ioannidis and Trikalinos (2007)). For such scenarios in which the between-study heterogeneity of a set of studies is expected to be at least considerable, another test was developed based on the principle of the statistical power (Schimmack (2012)). Following probability theory (Cohen (2013)), the Incredibility index test compares the expected number of statistically significant studies regarding the median observed statistical power of the studies. Again, if the number of expected statistically significant studies is sensibly lower than the actual number observed, it may be an indication of the presence of outcome reporting bias.

All this literature provides techniques and methodologies to detect the potential presence of publication and other reporting biases in a set of studies. Nevertheless, the impact of the presence of such biases also needs to be assessed to account for it and, if possible, measured to try to correct it. Several studies confirm the presumed effect of the publication bias on the results of summarizing analyses (like literature reviews, or meta-analyses, which will be introduced in detail later in the introduction of this thesis). A study re-analyzed 42 meta-analyses of clinical trials for nine drugs, including the unpublished data gathered from FDA trial data (Hart, Lundh, and Bero (2012)). After adding the unpublished data, the results changed sensibly and showed lower efficacy of the drug for 46% of the meta-analyses, no difference for 7%, and higher efficacy of the drugs for 46%. Moreover, the results for the single harm outcome after including the unpublished data showed more harm from the usage of the drug.

The outcome reporting bias may have an important impact on the results of summarizing analyses that should be assessed. The impact that selective outcome had on meta-analysis included in some of the 283 Cochrane reviews of clinical trials was also analyzed in the study (Kirkham et al. (2010)). The authors adjusted for outcome reporting bias by asking the original authors for the non reported data and including them on the analyses, and checked the differences in the 42 meta-analyses that originally had statistically significant results. After the correction, 19% of them became non-statistically significant, and 26% of them would have overestimated the treatment effect by 20% or more.

The concern about the importance of facing publication bias effects on the consolidation of new scientific knowledge is also present in the scientific editorial field, and there are actions taken to address this problem. One such being the growing acceptance of negative results (*The Missing Pieces: A Collection of Negative, Null and Inconclusive*

Results (n.d.), *AJG “The Negative Issue” November 2016 - American College of Gastroenterology* (n.d.), *General Information — eNeuro* (n.d.)), or the appearance of new journals specifically dedicated to negative results (*Negative Results — Home page of Scientific Journal Negative Results* (n.d.)). PLoS One journal has emphasized its policy of welcoming negative results publications, and in 2015 it even started a published collection of negative, null and inconclusive results (non-positive results) (*The Missing Pieces: A Collection of Negative, Null and Inconclusive Results* (n.d.)). Other journals are also opening sections dedicated to negative results (Dirnagl and Lauritzen (2010)). Nevertheless, what is called the file drawer effect, the imbalance between the number of publications with positive results and the number of publications with non-positive results, remains substantial (Kicinski, Springate, and Kontopantelis (2015)).

1.4 How the replication crisis, publication biases, and other issues affect neuroimage studies

The impact of the replication crisis and the low levels of replicability also reaches the neuroimaging research area (Nee (2019), B. O. Turner, Paul, Miller, and Barbey (2018)). The Nature journal published an editorial in 2017 exposing the concern about the low reproducibility of fMRI studies (“Fostering reproducible fMRI research” (2017), Bennett and Miller (2010)) and announcing the commitment of the editorial team to promoting the reproducibility of the studies published in the journal. They developed “an MRI-specific module to complement the methods reporting checklist” (“Fostering reproducible fMRI research” (2017)) already requested at submission time. A similar editorial note was published in 2018 by the editorial team of the Journal of Neuroscience (Picciotto (2018)). Also concerned about the replicability challenge of neuroimage studies, the note highlighted the recommendation of preregistering as a powerful tool to improve replicability (Poldrack et al. (2017)), agreeing on the desirability of preregistration as a policy to counteract publication bias already expressed as early as in 1993 (Dickersin and Min (1993)). Moreover, the editorial team reinforced the requirement for “authors to report experimental design and stats analyses fully” (Picciotto (2018)). The importance to address the replicability challenge to continue with the research in the neuroimaging area has also been reflexed upon by the Organization for Human Brain Mapping (OHBM), and the commitment of the organization to developing guidelines and recommended practices within the OHBM framework to successfully meet the reproducibility challenge (*Mumford_Keep Calm and Scan On - organization for human brain mapping.pdf* (n.d.)).

Not any less, publication bias also affects neuroimage studies. Among the factors eliciting publication bias aforementioned (J. P. A. Ioannidis (2005)), most of them affect the area: neuroimaging studies commonly have small sample sizes, the effect sizes tend to be small, there is great flexibility in designs and analytical methods and the field is hot with many teams pursuing publication. The actual impact of publication bias in neuroimage studies has been analyzed in some works (Jennings and Van Horn (2012), David et al. (2013)). For example, a study found a negative correlation between the study size and effect size, implying the presence of publication bias (Jennings and Van Horn (2012)). Moreover, the expected number of non-positive small studies was not met, also indicating publication bias. Using Cohen's d as a metric for effect size the visual plots were not conclusive, but were the quantitative methods assessing publication bias (Egger regression test with $F=6.7$ and $p=0.01$, Macaskill's regression method with $F=12.07$ and $p=0.0009$, and the Trim and Fill method with both tails $R_0 > 3$).

Besides the replication crisis and the presence of publication bias, the area is sensibly affected by specific problems with special incidence in neuroimaging, like usually modest study sample sizes (Szucs and Ioannidis (2019), Kennedy (2018), Lorca-Puls et al. (2018)) and high methodological heterogeneity (Specht (2020), Shuter, Yeh, Graham, Au, and Wang (2008)).

The vast amount of works published monthly makes it harder and harder to even keep track of all the new findings. Translating such a volume of novel information into consolidated knowledge is a task nearly unbearable without proper assistance, especially while separating the wheat from the grain and discerning the robust findings among the weaker ones.

Neuroimaging studies require expensive equipment like MRI scanners, which means that available sample sizes are usually small. To overcome this size-related problem there are initiatives in place trying to obtain larger data sets. One strategy deployed is the creation of databases compiling datasets from different sites and making them available to research teams that would benefit from the aggregation of analogous data and, therefore, the increment of sample sizes (Toga (2002)). Besides the long-established and well known NeuroVault (Gorgolewski et al. (2015)) or BrainMap (brainmap.org — *Home* (n.d.)), at the time of writing this thesis, there are more than 100 identified databases of neuroscience data (Acar, Seurinck, Eickhoff, and Moerkerke (2018)), many of them containing neuroimaging data and studies (Eickhoff, Nichols, Van Horn, and Turner (2016), J. Turner, Eickhoff, and Nichols (2017)). Another similar strategy is the creation of research consortiums formed by several different research groups and organized by study topics. Each group contributes their data and expertise and joint

studies are developed taking advantage of the much larger datasets attained (S. M. Smith and Nichols (2018), Bearden and Thompson (2017), Glasser et al. (2016)).

Finally, experimental designs vary sensibly among studies, with decisions that greatly affect the outcome (Carp (2012)). There are also initiatives to overcome this methodological heterogeneity, like the ones that propose the creation of standardized data processing pipelines (Carp (2012)). Ideally, these specifications would be followed by most of the researchers in the way that they would have a methodological guide for each type of analysis. This would provide a more homogeneous analytical framework attaining better methodological synchrony while performing analogous experiments (Beisteiner, Pernet, and Stippich (2019)).

1.5 Meta-analyses are very powerful to confront these challenges

The scientific community has developed and used several different strategies and techniques to face these challenges amongst which I must highlight meta-analysis. Not only meta-analysis currently is a widely used technique to deal with replicability issues (Hedges and Schauer (2019)) and, even, to analyze replicability rates (Braver, Thoemmes, and Rosenthal (2014)), but also has become an indispensable tool in modern science and the policy of enforcing its usage is becoming transversal (Perrin (2014)). Major scientific journals are more and more encouraging the performing of a prospective meta-analysis prior to designing new experiments (Young and Horton (2005), Collins and Tabak (2014)), to optimally deploy efforts and steer scientific progress. But before continuing with the value of the meta-analysis techniques, I shall explain what meta-analyses consist of.

A meta-analysis is a statistical methodology that is used to aggregate into a summarizing result the results of different studies collected in a systematic process (Borenstein, Hedges, Higgins, and Rothstein (2009)). The term meta-analysis was coined by Gene V. Glass when he intended to scientifically support his personal experience on the benefits of psychotherapy (M. L. Smith and Glass (1977)). Driven by his fuss with Hans Eysenck, who claimed (Eysenck (1952)) that psychotherapy was ineffective arguing the lack of evidence otherwise, he and Mary Lee Smith dedicated two years to collect all the available studies about the effects of psychotherapy. From the more than a thousand studies obtained, they selected 375 that were statistically fit (were methodologically correct and published resulting statistics) and computed the effect size of the psychological treatment in each study. Not only did they find out that the mean effect size was evidence of the therapy benefits (0.68), but they also conducted several sub-analyses to

study the effects of different types of therapies. Glass presented the technique, alongside the obtained results, in a presidential address to the American Educational Research Association in San Francisco in 1976. What is now a prominent milestone of the statistical sciences was received with major acceptance and the meta-analysis became the valued technique among the scientific community it is today. Except for Eysenck who, as far as eighteen years later, published a work arguing the unfitness of meta-analyses as “a model for studying the diversity of fields for which it has been used” (Eysenck (1994)).

From the moment they were first used, meta-analyses have proven on uncountable occasions their value, in many different research areas (Cumming (2013)). I will explain two examples that display the contributions that meta-analyses brought to health sciences and the importance of using meta-analyses on evidence-based practices and, consequently, on neuroimaging studies.

First, a work published in 2005 performed a historical review of childcare bibliography about associations between infant sleeping positions and sudden infant death syndrome (SIDS) and a cumulative meta-analysis of studies analyzing the relation (Gilbert, Salanti, Harden, and See (2005)). In their review, they confirmed that for most of the twentieth century, until as far as the late 80s, the commonly recommended practice was to make child sleep on their front. However, their cumulative meta-analysis discovered that already by 1970 there was quite reasonable evidence that back sleeping was safer. The resulting evidence kept growing with the following studies incorporated into the cumulative meta-analysis, although the individual evidence quality of these additional studies remained humble until the late 80's. Taking into account the number of infant deaths imputed to SIDS occurred in Europe, the USA and Australia, the authors estimated that had a meta-analysis been conducted in 1970, as many as 50.000 infant deaths could have been averted.

Second, in 1992 a meta-analysis was conducted over the studies of the efficacy of oral beta-blockers for reducing the risk of mortality in heart-attack survivors and published between 1972 and 1988 (Antman, Lau, Kupelnick, Mosteller, and Chalmers (1992)). It proved that after the ninth study, published in 1982, there was enough evidence of the benefits of the treatment, and the use of beta-blockers should have been established as a common practice. Not having done it resulted in losses of patients' lives and, later in time, waste of resources on redundant studies that did not provide additional findings once the original results were already replicated and confirmed.

Probably having considered these two examples, among others specifically mentioned in their editorial comment, since august 2005 the Lancet Journal requires authors of clinical trials submitted to the journal the inclusion of a summary of previous literature related

to the study to be published (Young and Horton (2005)). A published meta-analysis that serves the purpose of locating the new evidence to be published into the context of the existing evidence is also strongly recommended. Not only so, but in absence of such meta-analysis, authors are encouraged to perform their meta-analysis.

It is clear, therefore, that meta-analyses can be a powerful tool that increases the statistical power of the results by expanding the sample size. Having in mind that small sample sizes are one of the main issues of neuroimage studies, that is a very interesting feature that meta-analyses can contribute to neuroimage science. Furthermore, they are also very useful for identifying contradictions among the findings of the included studies that cannot be attributed to sampling error, detecting potential publication biases and even identifying factors in the data moderating the results (like methodological differences in the studies, clinical factors, etc). Once more, these are the most constraining drawbacks of neuroimage studies that I presented and are usually closely related to replicability or publication bias problems of scientific publications.

All these capabilities make meta-analysis an invaluable asset in facing the problems to consolidate new knowledge and push forward the advance of modern science.

1.6 Meta-analysis in voxel-based neuroimaging

As I will present accurately later, the main objective of this thesis was to develop a new method to perform meta-analyses of voxel-based neuroimaging studies and maximize the robustness of the results provided by the method and the derived conclusions. But first, to fully comprehend the work presented herein, some basic knowledge of neuroimaging, as well as some previous considerations regarding the application of meta-analysis in this discipline, should be presented.

Neuroimaging is the usage of one or various techniques to capture the image of the structure or functionality of the nervous system. Among all the different modalities of neuroimaging, I will cover the imaging of the brain using magnetic resonance imaging (MRI) or using positron emission tomography (PET). I will not get into the physics of how images are acquired, for the sake of this thesis will suffice to know that the result of both techniques is a 3D image divided into parcels of information. Such parcels are called voxels and contain the unit of information, either structural information (like volume) or functional information (like activation). Voxels can have isotropic dimensions (all sides equally-sized) or anisotropic (a different size for each side), and different sizes. Therefore, the total amount of voxels in a neuroimage varies considerably. As a reference, a typical

pre-processed structural MRI image may have isotropic 2mm-side voxels, which would add to around 1×10^6 voxels in an image.

Some challenges need to be faced when applying meta-analysis on neuroimaging studies, as standard statistics are not straightforwardly applicable due to the special constraints of neuroimaging data and publications. On the one hand, the main problems usually found while meta-analyzing the results of a systematic review on a neuroimaging topic are derived from the process of thresholding the results and reporting the summary. This process consists of first, with the complete brain map with all the voxels containing each the statistics, we apply a significance threshold by which they select only those voxels with values more statistically significant than the threshold and discard the rest. Each group of adjacent voxels in the remaining map will be named a cluster. Second, from each cluster, the authors get the voxel with maximum significance, which they will label as the peak of the cluster. At the end of this process, from an initial complete brain map of results they only report a list of peaks, containing the coordinates and the statistic of each peak. This process was developed due to the historical inability of distributing the amount of data required to publish full statistical maps, publishing the summarizing list of peaks once thresholded by a desired statistical value instead.

Nowadays, it is still not uncommon for the researchers to publish only the coordinates of the peaks of the clusters of statistically significant results of the studies, leaving the rest of the brain results unpublished. That implies that if we want to meta-analyze these areas and include the studies that did not report complete results, we need to impute the unpublished results for these areas and these studies.

On the other hand, another special constraint of voxel-based neuroimaging meta-analysis is the high dimensionality of neuroimages. As said before, commonly used voxel sizes result in images with dimensionality in the order of millions of data. This implies that we need to deal with a massive multiple comparison problem while estimating the significance of the results of our meta-analyses (Ashby (2019)).

1.7 Historic of methods and their problems

Originally, were two main types of meta-analysis methods for neuroimaging studies, regarding which data they could use. The image-based meta-analysis methods could be used only when all the complete statistical maps (hence image-based) of all the included studies were available. The coordinate-based meta-analysis methods could perform meta-analyses of neuroimaging studies that only reported coordinate-based results. The former consisted of applying standard statistics for most of the operations.

However, since the percentage of occasions on which full statistical maps are available is relatively low, image-based meta-analysis methods were seldomly applicable. To be able to include all those neuroimaging studies that reported only coordinate-based results, several coordinate-based meta-analysis methodologies have been developed over the last lustrum (Samartsidis, Montagna, Johnson, and Nichols (2017)). Two different approaches have been taken hitherto: kernel-based and model-based.

1.7.1 Kernel-Based

Kernel-based were the first methodologies developed, and the ones that are still commonly used in most published neuroimaging meta-analyses (Radua et al. (2014), Costafreda, David, and Brammer (2009), Turkeltaub et al. (2012), Wager, Lindquist, and Kaplan (2007)). The shared idea behind these methodologies is to first smooth the available data (reported cluster peaks) using a spatial kernel. Cluster peaks from each study are smoothed and combined into a corresponding study image. These study images are then combined to obtain the aggregated result, and the significance of the resulting image is estimated. Although it is out of the scope of this thesis to exhaustively review how the current methodologies perform these steps, I will peek at some details needed to understand the objectives of this thesis. First, there are different approaches to apply the kernel to the available peak-coordinates. The multi-kernel density analysis (MKDA) (Wager et al. (2007)) method generates a sphere with the coordinates of each peak as a center and fixed radius. All the voxels within this sphere are given a value of 1, and when combining different peak images with the corresponding spheres into a study image by just adding the images the voxels values are limited at 1 (if there are overlapping spheres the voxels keep their value at 1). Activation likelihood estimation (ALE) (Turkeltaub et al. (2012)) tries to model the probability that the activation was placed on the reported coordinate of the peak and its surrounding voxels. It models it by applying a Gaussian kernel with the center in the reported coordinate and enforcing that the sum of the voxels covered by the kernel adds to 1. Therefore, voxels closer to the peak have higher probabilities than voxels closer to the tails of the kernel. To combine different peak kernels into a study image, it uses the union of the probabilities given by kernelled peaks. Anisotropic effect-size seed-based d mapping (AES-SDM) (Radua et al. (2014)) takes a similar approach to ALE but tries to reproduce the original study maps. For each reported peak, it takes (when available) the effect size of the peak and applies an anisotropic Gaussian kernel. The kernels are dimensioned accordingly to a correlation template, trying to maintain, therefore, the spatial covariance of the brain tissue. To combine different peaks into a study image, a weighted mean of the peak images is computed.

To aggregate the generated study images, MKDA and AES-SDM compute the mean of the study images weighted by the study size (number of participants) and precision (only AES-SDM), AES-SDM uses signed effect sizes (e.g. positive for activations and negative for deactivations) and thus contradictory results will cancel.

A common trait of these methodologies is that, for the last step of estimating the significance image, they test for spatial convergence of peaks, applying a spatial randomization test that relies on spatial assumptions that sometimes are not met. A detailed explanation of this issue is explained in Chapter 3. but briefly, it may make the results either conservative or liberal, and it decreases the statistical power when there are several brain regions with relevant effects.

There was a kernel-based methodology that did not test for spatial convergence (Costafreda et al. (2009)). Like MKDA in generating a sphere from the reported peaks, for the estimation of the significance image, this method derives a parametric significance test based on the properties of the spatial Poisson process and then corrects for multiple testing using the false-discovery rate.

None of these methods could combine reporting cluster peaks with studies reporting full statistical maps, except for AES-SDM. Moreover, none of them took into consideration the non-reported results: the rest of the original study maps in which the data was discarded on the thresholding step of the original publications. Only AES-SDM tried to reproduce, to some extent, the non published original maps to be able to combine them with full statistical maps from studies publishing them.

1.7.2 Model-based

Another and newer approach to perform a meta-analysis of coordinate-based neuroimaging results are the so-called model-based methodologies. Based on spatial statistics, the principle behind them is to model the generation of peaks using stochastic processes. There have been several implementations published (Kang, Johnson, Nichols, and Wager (2011), Yue, Lindquist, and Loh (2012), Kang, Nichols, Wager, and Johnson (2014), Montagna, Wager, Barrett, Johnson, and Nichols (2018)), each taking different assumptions when determining the independence between sources for different tasks, or the relation between activations reported in the same study. Although such methodologies offer some interesting features, like providing a confidence level on the location of the activations, their practical usage is yet limited. At the time of writing this thesis, model-based methodologies had mostly been used to perform meta-analyses of emotional studies, where their ability to model different sources of activations for different

emotions had been particularly attractive. Like kernel-based methodologies, significance maps are estimated by using Monte Carlo testing schemas.

1.8 AES-SDM as our starting point

Among the different kernel-based methodologies available at the time of starting this thesis, we took AES-SDM as the starting point of our work. Already used profusely in many published meta-analyses, AES-SDM presents some specific and important strengths to deal with the exposed problems of neuroimaging studies and some aspects that could be improved. On the one hand, the capability of combining image-based and coordinate-based studies in a single meta-analysis is very powerful to augment the sample size supporting the evidence of meta-analysis results, as well as for the sake of expanding the exhaustive inclusion of studies in the analysis. SDM also offers tools to try to detect possible publication bias and to control unexplained heterogeneity. On the other hand, the recreation of non-published studies from lists of published coordinate-based results could be sensibly improved, as well as the estimation of the statistical significance maps of the results. As explained before, the spatial randomization test used to perform such estimation, also applied in most of the other methodologies currently used, can provide inaccurate results in some scenarios.

Chapter 2

Objectives

The objective of this thesis was to develop a method for the meta-analysis of neuroimaging studies that overcomes the drawbacks of the neuroimaging meta-analytic methods described above. To this end, we modified AES-SDM ([Radua et al. \(2014\)](#)) to incorporate several major changes that address each of the above limitations. This has resulted in a new method, called SDM-PSI, with profound differences from any previous neuroimaging meta-analytic method.

I divided this overarching objective into the following individual objectives:

1. Accurately identify and circumscribe the drawbacks of the neuroimaging meta-analytic methods present at the time of the conception of this thesis.
2. Develop the statistical techniques of the new method. Optimize the usage of maximum likelihood estimation (MLE) to estimate the distribution of the missing information of the neuroimaging studies and combine it with the usage of multiple imputation techniques to recreate this non-published information. This combination shall intend to obtain better estimates of effect sizes in areas where CBM studies did not report results.
3. Develop an algorithm for the new method viable for being implemented efficiently, providing an implementation that can be successfully used in real-world scenarios.
4. Modify the SDM meta-analytic method to incorporate these new statistical techniques, as well as other novel changes in the field such as a standard subject-based permutation test instead of the spatial permutation approach used by previous methods.

5. Design and develop an algorithm for the usage of the standard permutation test as part of the SDM method. A computationally efficient implementation of such an algorithm shall also be feasible.
6. Adopt the scope not only of theoretical-methodological work but also of practical work with usable results and implementations ready to be used by the neuroimage meta-analysis community. This shall include the development of powerful software and a graphical user interface.
7. Disseminate the new method to ease the adoption of its application by the scientific community. This included the release of the new software in a user-friendly form, writing and recording a visual publication [117], and proactively seeking collaboration studies with scientific teams from other institutions and labs that would provide the first publications making use of the new method.

The first objective is addressed in the first published article that comprises this thesis; the second and third objectives are addressed in the second published article; the fourth, fifth and sixth objectives are addressed in the third published article, and the seventh objective is addressed in the fourth published article.

Chapter 3

Articles included in this thesis

There are four articles that have been derived from this thesis and that are included in this chapter:

- Albajes-Eizagirre A, Radua J. What do results from coordinate-based meta-analyses tell us? *Neuroimage*. 2018 Aug 1;176:550-553. doi: 10.1016/j.neuroimage.2018.04.065. Epub 2018 May 3. PMID: 29729389.

JCR Impact factor: 5.812, First decile in Neuroimaging

- Albajes-Eizagirre A, Solanes A, Radua J. Meta-analysis of non-statistically significant unreported effects. *Stat Methods Med Res*. 2019 Dec;28(12):3741-3754. doi: 10.1177/0962280218811349. Epub 2018 Dec 4. PMID: 30514161.

JCR Impact factor: 2.291, First quartile in Statistics & Probability

- Albajes-Eizagirre A, Solanes A, Vieta E, Pomarol-Clotet E, Salvador R, Radua J. Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM. *Neuroimage*. 2019 Feb 1;186:174-184. doi: 10.1016/j.neuroimage.2018.10.077. Epub 2018 Oct 30. PMID: 30389629.

JCR Impact factor: 5.902, First decile in Neuroimaging

Corrigendum As of Dec 21th 2018, an authorship corrigendum was approved by Prof. Breakspear, Editor-in-Chief of Neuroimage. The authors Edith Pomarol-Clotet and Raymond Salvador, both with affiliations *FIDMAG Germanes Hospitalàries, Sant Boi de Llobregat, Barcelona, Spain* and *Mental Health Research Networking Center (CIBERSAM), Madrid, Spain*, were rightfully added.

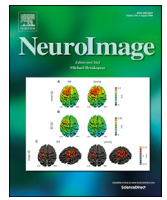
- Albajes-Eizagirre A, Solanes A, Fullana MA, Ioannidis JPA, Fusar-Poli P, Torrent C, Solé B, Bonnín CM, Vieta E, Mataix-Cols D, Radua J. Meta-analysis of Voxel-Based Neuroimaging Studies using Seed-based Mapping with Permutation of Subject Images (SDM-PSI). *J Vis Exp.* 2019 Nov 27;(153). doi: 10.3791/59841. PMID: 31840658.

JCR Impact factor: 1.163, third quartile in Multidisciplinary Sciences



Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

What do results from coordinate-based meta-analyses tell us?

Anton Albajes-Eizagirre^{a,b}, Joaquim Radua^{a,b,c,d,*}

^a FIDMAG Germanes Hospitalàries, Sant Boi de Llobregat, Barcelona, Spain

^b Mental Health Research Networking Center (CIBERSAM), Madrid, Spain

^c Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

^d Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

ARTICLE INFO

Keywords

Coordinate-based meta-analysis
Tests for spatial convergence
Familywise error rate
Activation likelihood estimation
Seed-based d mapping
Signed differential mapping

ABSTRACT

Coordinate-based meta-analyses (CBMA) methods, such as Activation Likelihood Estimation (ALE) and Seed-based d Mapping (SDM), have become an invaluable tool for summarizing the findings of voxel-based neuroimaging studies. However, the progressive sophistication of these methods may have concealed two particularities of their statistical tests. Common univariate voxelwise tests (such as the *t/z*-tests used in SPM and FSL) detect voxels that activate, or voxels that show differences between groups. Conversely, the tests conducted in CBMA test for “spatial convergence” of findings, i.e., they detect regions where studies report “more peaks than in most regions”, regions that activate “more than most regions do”, or regions that show “larger differences between groups than most regions do”. The first particularity is that these tests rely on two spatial assumptions (voxels are independent and have the same probability to have a “false” peak), whose violation may make their results either conservative or liberal, though fortunately current versions of ALE, SDM and some other methods consider these assumptions. The second particularity is that the use of these tests involves an important paradox: the statistical power to detect a given effect is higher if there are no other effects in the brain, whereas lower in presence of multiple effects.

Introduction

The exponential increase of voxel-based neuroimaging studies led to the need of methods that could summarize their results. Neuroimaging papers usually only report coordinates and statistics of the peaks (or “foci”) of the clusters of statistical significant voxels, and thus data extracted from these studies are a series of numeric tables that classical meta-analytic methods cannot combine. In this context, several developers introduced methods for conducting coordinate-based meta-analyses (CBMA), which are able to integrate this wealth of numeric information and return clear summary brain maps, thus shedding light on the neural substrates of many brain functions and neuropsychiatric disorders. Examples of these methods are Activation Likelihood Estimation (ALE) and Seed-based d Mapping (SDM), among others (Turkeltaub et al., 2002, 2012; Eickhoff et al., 2009, 2012; Laird et al., 2005; Radua et al., 2012, 2014; Radua and Mataix-Cols, 2009, 2012; Wager et al., 2007; Costafreda et al., 2009; Kang et al., 2011, 2014; Yue et al., 2012; Montagna et al., 2017; Tench et al., 2017).

Importantly, the statistical tests used by these methods have two

particularities as compared to the common univariate voxelwise tests present in neuroimaging software such as SPM or FSL. Univariate voxelwise tests, which may be used at subject-level, group-level or even study-level (e.g., to conduct a classic meta-analysis when all study data are available), assess whether a voxel shows not-null activation (i.e. blood oxygenation level dependent –BOLD– response is not zero), or whether a voxel shows not-null differences between groups (i.e. values in the two groups are not identical). Their statistics (usually *t/z*-values) summarize evidence against the null hypotheses “absence of BOLD response” or “absence of differences between groups”. Conversely, the tests conducted in CBMA test for “spatial convergence” of findings, i.e. they assess whether studies report more findings in the neighborhood of a given voxel than in the neighborhood of most voxels (Radua and Mataix-Cols, 2009). We show here that these tests rely on two spatial assumptions, whose violation may make their results either conservative or liberal, and that their statistical power decreases when there are multiple effects in the brain. We first present a toy meta-analysis to help us illustrate these points.

* Corresponding author. King's College London, Institute of Psychiatry, Psychology and Neuroscience PO 69, Division of Psychosis Studies, 16 De Crespigny Park, London, SE5 8AF, UK.

E-mail address: joaquim.radua@kcl.ac.uk (J. Radua).

<https://doi.org/10.1016/j.neuroimage.2018.04.065>

Received 15 January 2018; Received in revised form 27 April 2018; Accepted 28 April 2018

Available online 3 May 2018

1053-8119/© 2018 Elsevier Inc. All rights reserved.

A toy meta-analysis

For simplicity, we may imagine that the gray matter mask is composed of several independent voxels. The values of these voxels may be random t-values converted into effect sizes (Radua et al., 2012). Voxels whose values reach a given threshold may be considered “peaks” and set to “one”, whereas the value of the remaining voxels may be set to “zero”. The toy meta-analysis may simply consist of calculating the mean of the studies, separately for each voxel.

A test for spatial convergence could consist of repeatedly permuting the values “between the voxels” (i.e. randomizing the location of the peaks), to simulate meta-analyses in which any spatial convergence is only due to chance. The means of these permuted data would compose a null distribution from which we can derive the probability to obtain means as high as the original means by chance (i.e. the p-values). Specifically, each permutation would include the following steps: a) randomly swapping the effect-sizes between the voxels separately for each study; b) recalculating the means of the permuted studies, separately for each voxel; and c) saving the maximum of the means. If we aimed to control the familywise error rate (FWER) at 5% (i.e., to have a probability of 5% of making one or more type I errors), we would consider a voxel statistically significant if its original mean was higher than 95% of these maxima. The reader may see Fig. 1A and run Simulation 1 (Supplement) for an example. Note that for simplicity, the figure includes only six voxels, but the reader may set as many hundreds or thousands of voxels as desired in the simulations.

The reason why this test only saves the maxima is not related to CBMA but to the correction for multiple comparisons (Nichols and Holmes, 2002). It is obvious that 5% of the meta-analyses simulated in the permutations would have maxima that are higher than 95% of the maxima, and as we would wrongly consider these meta-analyses statistically significant, the FWER would be 5% (as we wish). The choice of FWER = 5% is arbitrary, other significance levels may be used.

Conversely, we would ask the reader to focus on how the test conducts a permutation: the null hypothesis is that peaks are randomly located within the gray matter mask, and to this end, we randomly reallocate the peaks during the permutations.

This procedure is radically different from the voxelwise permutations tests, such as those used in FSL “randomize”, which do not swap voxel values (Winkler et al., 2014). In a one-sample permutation, these tests multiply a random set of the individual images by -1 , and in a two-sample permutation, they randomly reassign the individuals to the two samples (Winkler et al., 2014). Afterwards, they recalculate the test statistics (e.g. t/z-values). For instance, to infer whether there are brain activation differences between males and females, these tests would first calculate the t-value image of the comparison between our sample of males (e.g., David, Ot and Nikolas) and our sample of females (e.g., Tina, Aina and Danai). In the first permutation, the tests could randomly assign Aina, David and Danai to the “male” group, and Ot, Tina and Nikolas to the “female group”, and they would re-calculate the t-value image. In the second permutation, they could randomly reassign Danai, Ot and David to the “male” group, and Nikolas, Tina and Aina to the “female group”, and they would re-calculate the t-value image again. And so on. With these random reassignments, the permutation tests break any association between brain activation and group labels (“male” and “female” in this example).

Spatial assumptions

In the unrealistic toy meta-analysis, the permutation test was accurate because i) each voxel was independent from its neighbors, and ii) each voxel had the same probability to have a “false” peak. However, the data may not always meet these assumptions, as detailed in the following.

First, in real gray matter, voxels correlate with their neighbors,

local peaks from the same cluster tend to be very close, peaks from neighboring related clusters are closer than peaks from independent clusters, and etcetera. If the data simulated in the permutations do not have the spatial structure of the original data, there are differences between the original data and the permuted data that are unrelated to spatial convergence but due to the differences in spatial structures. For instance, two local peaks from the same cluster are usually very close, whereas in the permutations they could be at any distance. The reader may run Simulation 2 (Supplement) for an example where the original data do not meet the assumption of spatial independence because the peaks are very close, simulating close local peaks from the same cluster. In this example, the test would be substantially conservative.

Moreover, the destruction of the spatial structure in the permutations invalidates the use of spatial statistics, in which p-values are derived from the cluster sizes (or similar measures such as cluster masses or TFCEs (Smith and Nichols, 2009)). The spatial correlation between neighbor voxels involves that statistically significant voxels tend to be together forming clusters. If these correlations are not present in the permutations, statistically significant voxels are sparse and do not form clusters during the permutations, inflating the statistical significance (the p-values associated to the different cluster sizes become too small, i.e., clusters as large as the ones observed in the unpermuted data would be extremely unlikely in the permuted data).

To preserve the spatial structure, permutation tests should ensure that effect sizes from neighbor voxels remain together, or that the Euclidean distances between peaks are unmodified, in all permutations. Multi-level Kernel Density Analysis (MKDA) introduced the swapping of blocks of voxels, rather than voxels, in an attempt to preserve this spatial structure (Wager et al., 2007), and similar approaches were subsequently added to ALE (Eickhoff et al., 2009) and SDM (Radua et al., 2012).

Second, probably all studies cover voxels that are mostly composed of gray matter, but only some of the studies may cover voxels that are only marginally composed of gray matter. If only some of the studies cover a voxel, it is obviously less likely that a study reports a peak in this voxel, violating the assumption of homogeneity of the probability to have a “false” peak. The reader may run Simulation 3 (Supplement) for an example where this violation would make the test substantially liberal. Fortunately, some modern CBMA methods such as SDM include accurate tissue templates to minimize this effect (Radua et al., 2011; Peters et al., 2012).

These issues do not apply to the voxelwise permutations tests (Winkler et al., 2014) because they do not swap values between voxels.

Statistical power and number of effects

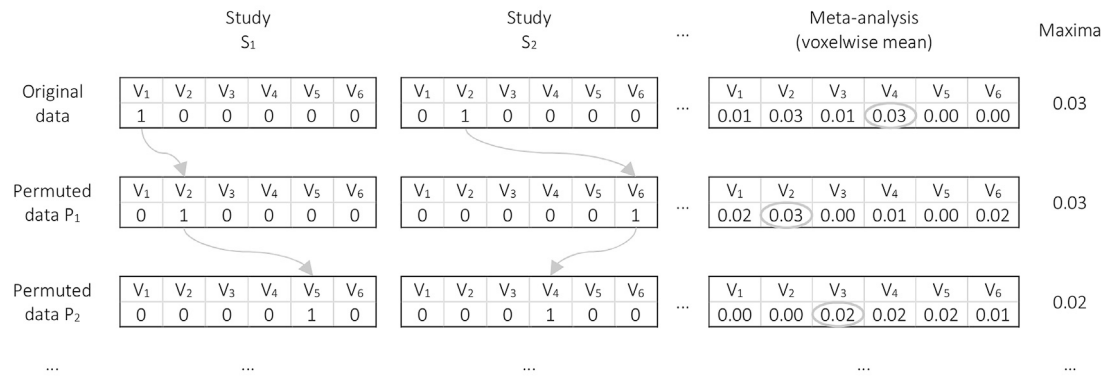
Imagine that all studies in the toy meta-analysis reported a peak in the first voxel and no other findings. In the permutations, each study would have one peak randomly located in any of the voxels, and thus the probability that a voxel of a study had a peak would be $1/N_{\text{voxels}}$. The probability that this voxel had a peak in all studies would be $1/N_{\text{voxels}}$ raised to N_{studies} . Finally, we could multiply this probability and the number of voxels to have the probability that any voxel had a random peak in all studies, i.e. the probability to have a meta-analytic value as high as the meta-analytic value of the first voxel in the original data:

$$P_1 = \left(\frac{1}{N_{\text{voxels}}}\right)^{N_{\text{studies}}} \cdot N_{\text{voxels}} = \left(\frac{1}{N_{\text{voxels}}}\right)^{N_{\text{studies}} - 1}$$

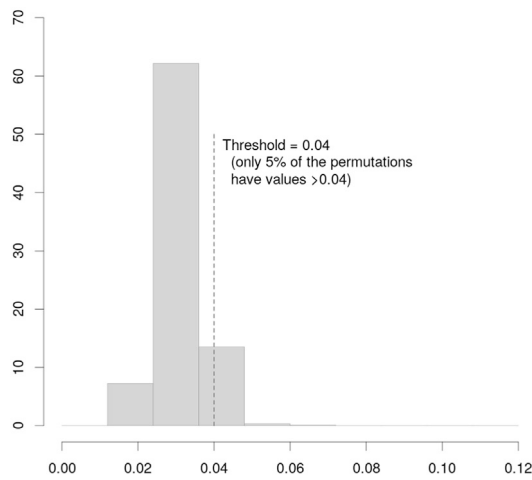
Now, imagine that in addition to the peak in the first voxel, all studies reported a peak in the second voxel. Following a similar argument, we could derive that the probability to obtain, in a permutation, a meta-analytic value as high as the meta-analytic value of the first two voxels in the original data would be:

A) A toy meta-analysis

Permutations:



Assessment of statistical significance:



As observed in the histogram of the maxima (left), 95% of the maxima are ≤ 0.04 .

Thus, only those voxels from the original meta-analysis with an effect size > 0.04 are statistical significant.

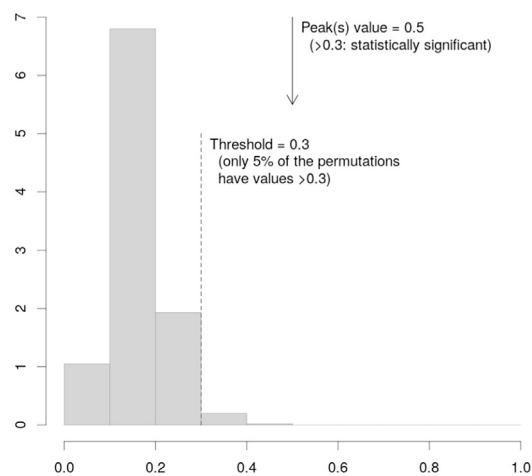
As shown above, none of the voxels from the original meta-analysis has an effect size > 0.04 :

V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
0.01	0.03	0.01	0.03	0.00	0.00

Therefore, there are no statistically significant findings.

B) Statistical power and number of effects

One true peak in 50% studies, no false peaks:



Three true peaks in 50% studies, no false peaks:

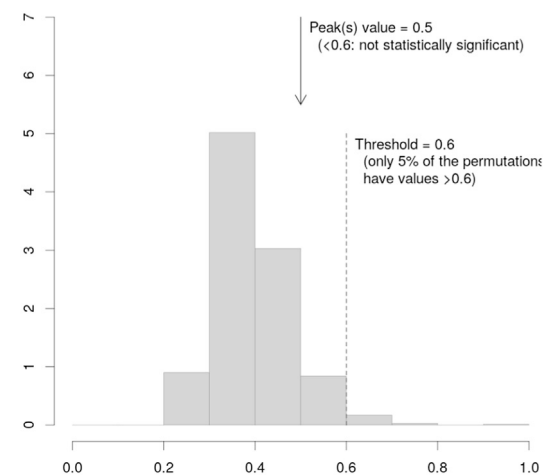


Fig. 1. A toy meta-analysis and the relationship between the number of effects and statistical power.

$$P_2 = \left(\frac{2}{N_{\text{voxels}}}\right)^{N_{\text{studies}}} \cdot N_{\text{voxels}} - \left(\frac{1}{\binom{N_{\text{voxels}}}{2}}\right)^{N_{\text{studies}}} \cdot \binom{N_{\text{voxels}}}{2} = \left[2^{N_{\text{studies}}-1} \cdot \left(2 - \frac{1}{(N_{\text{voxels}}-1)^{N_{\text{studies}}-1}}\right)\right] \cdot P_1$$

Given that the expression between square brackets is substantially larger than one, P_2 is substantially larger than P_1 , i.e. there would be a substantial increase in the probability that a voxel has a meta-analytic value as high as the meta-analytic values in the original data, and this increase involves a substantial decrease in statistical power. The user may run Simulations 3 and 4 (Supplement) for an example of substantially lower statistical power when the same test is conducted in the presence of multiple effects than when is conducted in the presence of a single effect. Interestingly, these simulations show that the reduction of power might be stronger when the number of voxels is larger.

Fig. 1B also shows how the threshold substantially increases (reducing statistical power) in the presence of multiple effects. In the left example, 50% of the studies have one true peak and no false peaks, and this true peak is statistically significant (its value, 0.5, is larger than the threshold, 0.3). In the right example, 50% of the studies have three true peaks and no false peaks, and these true peaks are not statistically significant (their value, 0.5, is lower than the threshold, 0.6).

The situation is different in univariate voxelwise tests, because the uncorrected p-value of one voxel does not depend on the values of the other voxels (beyond the correlation expected between correlated voxels, e.g. in real data adjacent voxels have similar p-values).

Are the tests comparable?

Whether these tests are comparable and to what extent, may be a matter of discussion. On the one hand, they usually have different uses, as group analyses of individual images use common voxelwise tests, while CBMA use tests for spatial convergence. However, this association might simply be due to the limited availability of data for voxel-based meta-analyses, as they can use common voxelwise tests if all study data are available. And vice versa: there is indeed no theoretical impediment to use a test for convergence to conduct a group-level test (i.e., looking for the convergence of the peaks found in the subject-level tests). On the other hand, common voxelwise tests have a set of steps that result in a threshold that ensures that, in the absence of true effects, there is only 5% probability that a voxel is statistically significant, and tests for spatial convergence have different steps and result in a different threshold that ensures that, if activations or differences are distributed uniformly throughout the gray matter, there is only 5% probability that a voxel is statistically significant. In other words, researchers can conclude that voxels with values above the voxelwise-test-threshold “activate” (or show differences between groups), whereas voxels with values above the test-for-convergence-threshold “activate more than most voxels do” (or show larger differences between groups than most voxels do). Therefore, the tests may have a common goal, but are used in different scenarios and assess indeed different things.

Conflicts of interest

The authors report no conflicts of interests related to this manuscript.

Acknowledgements

This work was supported by Miguel Servet Research Contract MS14/00041 to JR and Research Project PI14/00292 from the Plan Nacional de I+D+i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.


Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.neuroimage.2018.04.065>.

References

- Costafreda, S.G., David, A.S., Brammer, M.J., 2009. A parametric approach to voxel-based meta-analysis. *Neuroimage* 46 (1), 115–122.
- Eickhoff, S.B., et al., 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30 (9), 2907–2926.
- Eickhoff, S.B., et al., 2012. Activation likelihood estimation meta-analysis revisited. *Neuroimage* 59 (3), 2349–2361.
- Kang, J., et al., 2011. Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *J. Am. Stat. Assoc.* 106 (493), 124–134.
- Kang, J., et al., 2014. A Bayesian hierarchical spatial point process Model for multi-type neuroimaging meta-analysis. *Ann. Appl. Stat.* 8 (3), 1800–1824.
- Laird, A.R., et al., 2005. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25 (1), 155–164.
- Montagna, S., et al., 2017. Spatial Bayesian latent factor regression modeling of coordinate-based meta-analysis data. *Biometrics*.
- Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15 (1), 1–25.
- Peters, B.D., et al., 2012. White matter development in adolescence: diffusion tensor imaging and meta-analytic results. *Schizophr. Bull.* 38 (6), 1308–1317.
- Radua, J., Mataix-Cols, D., 2009. Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *Br. J. Psychiatry* 195 (5), 393–402.
- Radua, J., Mataix-Cols, D., 2012. Meta-analytic methods for neuroimaging data explained. *Biol. Mood Anxiety Disord.* 2, 6.
- Radua, J., et al., 2011. Voxel-based meta-analysis of regional white-matter volume differences in autism spectrum disorder versus healthy controls. *Psychol. Med.* 41 (7), 1539–1550.
- Radua, J., et al., 2012. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *Eur. Psychiatry* 27 (8), 605–611.
- Radua, J., et al., 2014. Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Front. Psychiatry* 5, 13.
- Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44 (1), 83–98.
- Tench, C.R., et al., 2017. Coordinate based random effect size meta-analysis of neuroimaging studies. *Neuroimage* 153, 293–306.
- Turkeltaub, P.E., et al., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16 (3 Pt 1), 765–780.
- Turkeltaub, P.E., et al., 2012. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum. Brain Mapp.* 33 (1), 1–13.
- Wager, T.D., Lindquist, M., Kaplan, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect Neurosci.* 2 (2), 150–158.
- Winkler, A.M., et al., 2014. Permutation inference for the general linear model. *Neuroimage* 92, 381–397.
- Yue, Y.R., Lindquist, M.A., Loh, J.M., 2012. Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *Ann. Appl. Stat.* 6 (2), 697–718.

Meta-analysis of non-statistically significant unreported effects

Anton Albajes-Eizagirre,^{1,2} Aleix Solanes^{1,2,3}
and Joaquim Radua^{1,2,3,4,5} 

Statistical Methods in Medical Research
0(0) 1–14

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0962280218811349

journals.sagepub.com/home/smm



Abstract

Published studies in Medicine (and virtually any other discipline) sometimes report that a difference or correlation did not reach statistical significance but do not report its effect size or any statistic from which the latter may be derived. Unfortunately, meta-analysts should not exclude these studies because their exclusion would bias the meta-analytic outcome, but also they cannot be included as null effect sizes because this strategy is also associated to bias. To overcome this problem, we have developed MetaNSUE, a novel method based on multiple imputations of the censored information. We also provide an R package and an easy-to-use Graphical User Interface for non-R meta-analysts.

Keywords

Meta-analysis, non-statistically significant unreported effects, interval censoring, multiple imputation

1 Introduction

When conducting a meta-analysis, analysts may find that some studies report that the effect (e.g. a difference between two groups) was not statistically significant but do not report the specific value of the outcome (e.g. a change in blood pressure) or any statistic from which the outcome may be derived (e.g. a *t*-value). These non-statistically significant unreported effects (NSUEs) should neither be excluded from the meta-analysis nor be included assuming them to be null, as we expose hereafter. On the one hand, note that they probably correspond to small effect sizes, and thus their exclusion would inflate the meta-analytic effect size: we would include all studies that find larger changes in blood pressure, while we would exclude (some or all) studies that find smaller changes. But on the other hand, the effect size of the latter studies is small but probably not null; including them as null effect sizes would shrink the meta-analytic effect size toward zero.¹

To include correctly studies with NSUEs into meta-analyses, we developed MetaNSUE, a novel method based on multiple imputation algorithms. Briefly, this novel method consists in obtaining maximum likelihood estimations (MLEs) of the censored outcomes (i.e. the NSUEs), adding noise to these estimations to multiply impute the NSUEs, conducting a random-effects meta-analysis for each set of imputations, and finally combining these meta-analyses. The reason to use a multiple imputation approach, rather than directly using the MLEs of the parameters of the meta-analysis, is that the MLE tends to underestimate the between-study heterogeneity.²

We want to note that we already published and validated a preliminary version of the method.³ In its empirical validation, we observed inflated estimations when conducting a standard meta-analysis without NSUEs, estimations shrunk toward zero when converting NSUEs to zeros, and nearly unbiased estimations when using

¹FIDMAG Germanes Hospitalàries, Barcelona, Spain

²Mental Health Research Networking Center (CIBERSAM), Madrid, Spain

³Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

⁴Department of Clinical Neuroscience, Centre for Psychiatric Research and Education, Karolinska Institutet, Stockholm, Sweden

⁵Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

Corresponding author:

Joaquim Radua, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London SE5 8AF, UK.

Email: joaquim.radua@kcl.ac.uk

MetaNSUE. However, we later detected that the preliminary version of the method may be unstable in two extreme scenarios, namely when only few studies of the meta-analysis report the specific value of the outcome and in meta-analyses in which studies use unconventional bounds (e.g. $z \geq 10$).

In this paper, we present a robust, final version of the novel method for meta-analyses with NSUEs and its R package available at the CRAN repository. In the Supplement, we also describe its new Graphical User Interface (GUI), free to download from www.metansue.com.

1.1 Worked example

Let us imagine that we aim to conduct a meta-analysis of the drug-related change in blood pressure and that we can include 10 studies, of which four are NSUEs. In other words, we can find or derive the t -values for six of the studies, but we only know that the change did not reach statistical significance for the other four.

As the reader can check by running the code provided in the Supplementary Material, if we discard studies with NSUEs, the effect size would be 0.52, whereas if we assume that NSUEs were null, the effect size would be 0.32 (Table 1 and Figures 1 and 2).

It is impossible to know the true effect size of these made-up data, but it is clear that at least one of the two estimations is incorrect. According to the method presented in this paper, the estimated effect size would be 0.42, for which both earlier approaches would be incorrect. We will resume this example later.

1.2 Notation

In a standard meta-analysis, analysts first estimate the effect size (e.g. a standardized mean difference or an odds ratio) of each study, along with its variance. We will denote the effect size and variance of the i th study as y_i and v_i , respectively. They then commonly combine the effect sizes using a weighted linear model, in which each element of the diagonal of the weighting matrix W is the inverse of the variance of the corresponding study plus the heterogeneity. The latter refers to the variation between studies that it is not due to chance, and we denote it as τ^2 . Some meta-analysts assume that τ^2 is zero, conducting a so-called “fixed-effects” meta-analysis, but this assumption is generally discouraged.

The main analysis is usually a weighted mean, and thus the design matrix X is a column of ones. In meta-regressions, each modulator is an additional column of the design matrix. The results of the model are the weighted mean (or the coefficients of the model β) and its variance (or the variance-covariance matrix Λ)

$$\begin{aligned}\beta &= \Lambda \times X^T \times W \times Y \\ \Lambda &= (X^T \times W \times X)^{-1}\end{aligned}\tag{1}$$

To explore the heterogeneity, meta-analysts commonly conduct a test with the Cochran’s Q statistic, which follows a χ^2 distribution and report descriptive statistics such as the H^2 (ratio of total variability to sampling variability) and the I^2 (percentage of total variability due to heterogeneity)

$$\begin{aligned}H^2 &= 1 + \frac{\hat{\tau}^2}{df} \cdot \text{trace}(P_{FE}) \\ I^2 &= 1 - \frac{1}{H^2}, \quad I^2 \geq 0 \\ Q &= Y^T \times P_{FE} \times Y, \quad Q \geq 0\end{aligned}\tag{2}$$

Note that “trace” is the sum of the main diagonal elements, W_{FE} is the weighting matrix for a fixed effects model, and P_{FE} is

$$P_{FE} = W_{FE} - W_{FE} \times X \times (X^T \times W_{FE} \times X)^{-1} \times X^T \times W_{FE}\tag{3}$$

2 NSUE meta-analysis

The core steps of an NSUE meta-analysis are:

- (1) Preparing the data
- (2) Conducting an MLE of the unknown effect sizes and between-study heterogeneity³

Table 1. Results of the example meta-analysis.

		Discard NSUEs	Set NSUEs to zero	MetaNSUE
Coefficient	Estimate	0.52	0.32	0.42
	95% CI	0.36–0.69	0.15–0.49	0.27–0.57
	z	6.2	3.6	5.6
	p	<0.001	<0.001	<0.001
Heterogeneity	τ^2	0.0	0.031	0.0003
	I^2	1.00	1.69	1.01
	I^2	0%	41%	0.52%
	Q	0.9	15.3	2.65
	p	0.97	0.084	0.98

NSUEs: non-statistically significant unreported effects.

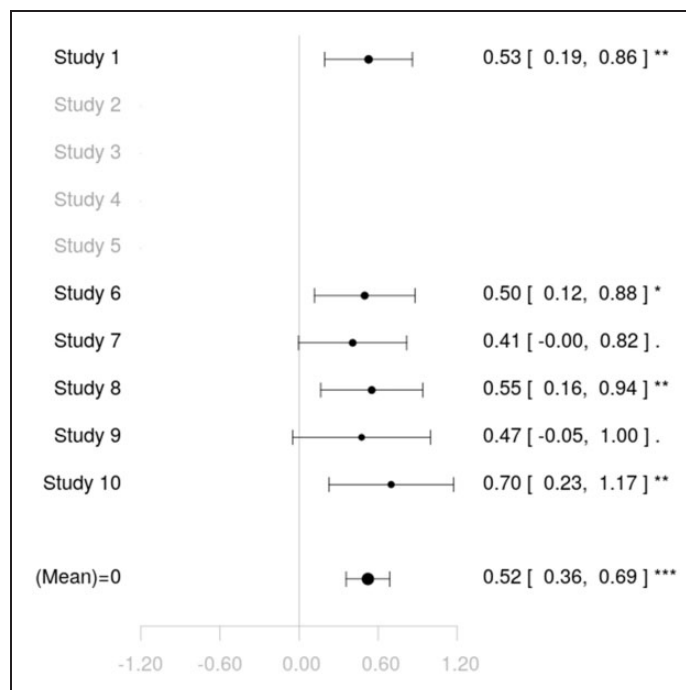


Figure 1. Forest plot of the example meta-analysis, discarding studies with non-statistically significant unreported effects (NSUEs). MetaNSUE marks with asterisks statistically significant results.

- (3) Imputing these unknown effect sizes based on the MLE⁴ many times
- (4) Meta-analysing each set of imputations separately⁵
- (5) Combining the results of the meta-analyses⁶
- (6) Hypothesis testing using the combined statistics⁷

MetaNSUE software automatically conducts steps 2 to 6 with the R function “meta.”

2.1 Preparing the data

Prior to the MLE and further steps, MetaNSUE meta-analysts must convert the known statistics (e.g. t -values) into effect sizes and include them in an object of class “nsue.” We provide three simple R functions for converting standardized mean changes, standardized mean differences and coefficients of correlations into effect sizes using standard formulas (see Supplementary Material) but also a generic function for converting other outcomes.

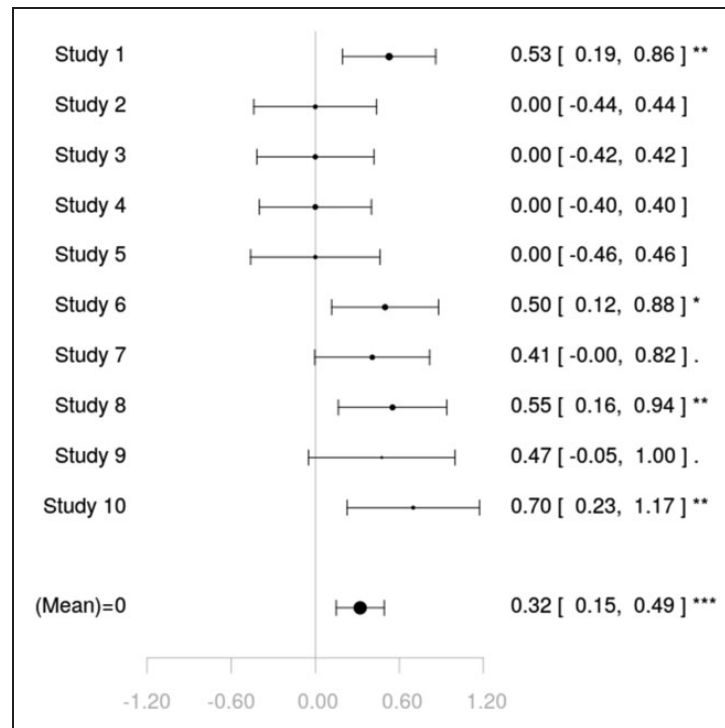


Figure 2. Forest plot of the example meta-analysis, assuming that non-statistically significant unreported effects (NSUEs) are null. MetaNSUE marks with asterisks statistically significant results.

Independent of the function used, MetaNSUE meta-analysts must code values from studies with NSUEs with “NA”. This way, the above functions know that these studies have NSUEs and use the significance level of the studies to calculate the effect sizes corresponding to the bounds of non-statistical significance (i.e. the unknown effect sizes must be within these bounds; otherwise, they would have reached statistical significance). Note that the significance level may be different for each study.

We must highlight that the variance of some measures of effect size depends on the effect size, but the latter are unknown in studies with NSUEs. To allow a straightforward estimation of the variance in the subsequent imputation steps, the objects of class “nsue” include a function to derive the variance from the effect size.

Finally, we must comment that MetaNSUE meta-analysts can include in an object of class “nsue” their own functions to derive the variance, to impute (useful for outcomes other than those already convertible with the functions provided with the software) and to back-transform (useful for transformed measures such as correlation Fisher’s z or log odds ratio).

2.2 MLE

We want to highlight that MetaNSUE conducts an MLE with the only aim of providing a basis for the subsequent multiple imputations of the unknown effects. It would be incorrect to understand the estimations resulting from this step as meta-analytic results because MLE tends to underestimate the between-study heterogeneity,² which is instead well estimated in the subsequent multiple imputations. We conducted simulations, analogous to the ones described in section 3, to show this effect: MLEs of heterogeneity were (nearly) unbiased in the absence of heterogeneity but shrunk toward zero (-0.04) in the presence of heterogeneity. However, the latter bias decreased to negligible levels in the subsequent multiple imputations (-0.01).

The reader may also wonder why we did not directly use restricted maximum likelihood (REML) in the MLE step. The reason is that, unfortunately, its use in this step would be problematic. REML does not use y but linear combinations of y , and any linear combination c that included at least one study with NSUE would become interval censored. For example, in a meta-analysis of two studies with known effects ($y_1 = 0.6$ and $y_2 = 0.8$) and one study with NSUE (y_3 between -0.5 and 0.5), the linear combination $c_1 = (2y_1 - y_2 - y_3)/3$ would not be an exact number but the interval $(-0.03, 0.3)$. Similarly, the linear combination $c_2 = (2y_2 - y_1 - y_3)/3$ would neither

be an exact number but the interval (0.17, 0.5). Thus, most (or all) “studies” of REML would be interval censored, amplifying the problem.

2.2.1 The log-likelihood

MLE attempts to find the parameters that maximize the likelihood of the data. Specifically, the likelihood to maximize is the multiplication of the likelihood of each reported effect size and the likelihood that each unreported effect size lies within its two effect size bounds, as in a Tobit regression for censored data.^{8,9} The likelihood of a reported effect size is the probability density function (pdf), and the likelihood that an unreported effect size lies within its two effect size bounds is the difference between the cumulative distribution function (cdf) evaluated at the upper size bound and the cdf evaluated at the lower size bound

$$\begin{aligned} L(\beta; y_i, \dots; v_i, \dots; y_{\alpha/2,j}, y_{1-\alpha/2,j}, \dots) \\ = \prod_{i=1}^{N_1} \text{pdf}(y_i; v_i | \beta) \cdot \prod_{j=1}^{N_2} (\text{cdf}(y_{1-\alpha/2,j}, v_{1-\alpha/2,j} | \beta) - \text{cdf}(y_{\alpha/2,j}, v_{\alpha/2,j} | \beta)) \end{aligned} \quad (4)$$

where N_1 is the number of reported effect sizes and N_2 is the number of studies with NSUEs.

The parameters to estimate include not only the coefficients of interest (usually the intercept, i.e. the main outcome of a meta-analysis) but also regressors that could improve the estimation of the effect size of a study. The preliminary version of MetaNSUE also included the between-study heterogeneity τ^2 as a parameter to estimate, but the final version conducts this estimation in a separate step: it first conservatively estimates the coefficients discarding extreme studies (see the next section) and afterwards estimates τ^2 including all studies. Remember in any case that MLEs represent only an intermediate step of the process, and the following steps of MetaNSUE include τ^2 in the estimations of the coefficients.

Assuming normality (e.g. when samples are sufficiently large), the pdf is simply

$$\text{pdf}(y_i, v_i | \beta) = \frac{1}{\sqrt{2\pi v_i}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(y_i - X_i \cdot \beta)^2}{v_i}\right) \quad (5)$$

Similarly, the cdf is

$$\text{cdf}(y_i, v_i | \beta) = \Phi\left(\frac{y_i - X_i \cdot \beta}{\sqrt{v_i}}\right) \quad (6)$$

where Φ is the cumulative distribution function of the standard normal distribution.

In practice, it is more convenient to minimize the minus log-likelihood than to directly maximize the likelihood. Thus, the expression to minimize is the sum of the minus logarithms of the pdf or the difference of cdf of each study.

The minus logarithm of the pdf is relatively simple

$$\begin{aligned} -\log(\text{pdf}(y_i, v_i | \beta)) &= \frac{1}{2} \cdot \log(2\pi) + \frac{1}{2} \log(v_i) + \frac{1}{2} \cdot \frac{(y_i - X_i \cdot \beta)^2}{v_i} \\ &= (ct.) + \frac{1}{2} \cdot \left[\log(v_i) + \frac{(y_i - X_i \cdot \beta)^2}{v_i} \right] \end{aligned} \quad (7)$$

Conversely, the logarithm of the difference of the cdf can be numerically unstable in extreme scenarios where the two expressions inside the cdf are large and close. To minimize numerical errors, the final version of MetaNSUE includes the following transformation

$$\log(\Phi(m) - \Phi(n)) = \Phi_{\log}(m) + \log(-\text{expml}(\Phi_{\log}(n) - \Phi_{\log}(m))) \quad (8)$$

where $\text{expml}(x)$ is equivalent to $e^x - 1$ but has increased numerical accuracy. For $m = 10$ and $n = 9.9$, the initial expression wrongly returns $-\infty$, whereas the last equivalent expression correctly returns -52.7 .

MetaNSUE internally uses the R functions “optimize” and “optim” to minimize the minus log-likelihood when there is a single coefficient and when there are two or more coefficients, respectively. It previously conducts a fixed-effects meta-analysis (assuming NSUEs to be the mean of their upper and lower bounds) to estimate the initial values of the parameters, required for the latter function.

2.2.2 Leave-one-out protection

In the validation of the preliminary version of MetaNSUE, we showed that this method returned less biased estimations than discarding studies with NSUEs or considering that NSUEs are null, even if nearly all studies were NSUEs. However, in the latter scenarios, there is still a possibility that a single or few studies with known measures drive the MLE, potentially leading to a falsely positive meta-analysis. MetaNSUE meta-analysts may detect this situation in a subsequent leave-one-out jack-knife analysis when a single study clearly produces the artifact. For example, this was the case for reward feedback in our previous meta-analysis of ventral striatal activation in psychosis.³ However, the issue could potentially be unnoticed in less extreme situations.

To minimize this possibility, the final version of MetaNSUE includes a leave-one-out protection within the MLE step, which consists in iteratively discarding the study that increases the most the MLE effect size. Please note that this protection only involves the MLE step and has no relationship with the leave-one-out sensitivity analyses conducted with the R function “leave1out” (which iteratively repeats the whole meta-analysis). Specifically, the method first estimates the MLE with all studies but the first, then with all studies but the second, then with all studies but the third, and so on, and only keeps the combination returning the lowest absolute MLE effect size.

When the number of studies is large, the software repeats this iteration several times, discarding one additional study in each iteration, until the probability that the meta-analysis is falsely positive due to the issue is not higher than 0.05 even in the worst-case scenario (i.e. when all studies with non-statistically significant effect sizes are NSUEs). Please see the Supplementary Material for details.

If the meta-analysts do not aim to conduct a simple, mean meta-analysis but a meta-regression or another linear model, MetaNSUE will discard the studies according to the effect size of the hypothesis of interest, rather than according to the effect size of the intercept. For this reason, MetaNSUE meta-analysts must specify the hypothesis to contrast before conducting the MLE, and must conduct a new MLE for each other new hypothesis.

2.2.3 Estimation of heterogeneity τ^2

Once the coefficients have been estimated, MetaNSUE conducts a second MLE step to estimate the between-study heterogeneity parameter τ^2 , this time using all studies. The likelihood to maximize is the same as in equation (4), but the pdf and cdf described in equations (5) and (6) include τ^2 , and use $e_i = y_i - X_i \cdot \beta$ instead of y_i .

2.3 Multiple imputation

The multiplication of X and β returns the expected value of the effect size of each NSUE study. For example, if $\beta = (3.2, 0.2)$ and $X_i = (1, 2)$, the expected value of the effect size for this i th study is $1 \times 3.2 + 2 \times 0.2 = 3.6$. Then, the multiple imputation step consists in adding Gaussian noise to the expected value in order to have realistic imputations. In the example, MetaNSUE would add random noise to 3.6 so that imputation 1 could be 3.3, imputation 2 could be 3.7, etc. Note that imputing the NSUE studies using their *expected* value without added noise would erroneously mean assuming that the variability is null.

The addition of noise must meet two conditions: (a) the noise must be the sum of the variance of the effect size plus τ^2 and (b) the noisy imputation must be between the effect size bounds. The truncated normal distribution accommodates these conditions, and thus each imputation is a random generation of

$$\text{pdf}(y_i|\theta) = \frac{\frac{1}{\sqrt{v_i + \tau^2}} \cdot \phi\left(\frac{y_i - \hat{y}_i}{\sqrt{v_i + \tau^2}}\right)}{\Phi\left(\frac{y_{i,1-\alpha/2} - \hat{y}_i}{\sqrt{v_i + \tau^2}}\right) - \Phi\left(\frac{y_{i,\alpha/2} - \hat{y}_i}{\sqrt{v_i + \tau^2}}\right)}, \quad y_i \in (y_{i,\alpha/2}, y_{i,1-\alpha/2}) \quad (9)$$

where ϕ is the probability density function of the standard normal distribution.

Unfortunately, for some measures such as standardized mean differences, the variance depends on the imputed effect sizes (rather than on the sample size alone). In those imputations in which the added noise decreases the effect size, the variance is smaller and the study receives more weight. Conversely, in those imputations in which the added noise increases the effect size, the variance is larger and the study receives less weight. This underweight implies that the study receives systematically more weight in the imputations that underestimate its effect size than

in the imputations that overestimate it, potentially shrinking the estimations combined using Rubin's rules toward zero. In order to avoid this bias, functions to impute these effect sizes might weight the probability of a given effect size by the inverse of the weight that the effect size would receive in a subsequent meta-analysis

$$\text{pdf}(y_i|\theta) = \frac{\left(\hat{v}_{i,y_i} + \hat{\tau}^2\right) \cdot \frac{1}{\sqrt{\hat{v}_{i,y_i} + \hat{\tau}^2}} \cdot \phi\left(\frac{y_i - \hat{y}_i}{\sqrt{\hat{v}_{i,y_i} + \hat{\tau}^2}}\right)}{\int_{x=y_{i\alpha/2}}^{y_{i1-\alpha/2}} \left(\left(\hat{v}_{i,x} + \hat{\tau}^2\right) \cdot \frac{1}{\sqrt{\hat{v}_{i,x} + \hat{\tau}^2}} \cdot \phi\left(\frac{x - \hat{y}_i}{\sqrt{\hat{v}_{i,x} + \hat{\tau}^2}}\right)\right)}, \quad y_i \in (y_{\alpha/2}, y_{1-\alpha/2}) \quad (10)$$

In practice, the software divides the continuum between the bounds in a number of bins, and randomly selects a bin according to its probability.

2.4 Meta-analysis

After the multiple imputation step, there are several datasets of complete data (i.e. without NSUEs), with each dataset including one (known or imputed) effect size per study.

These datasets can be meta-analyzed using standard formulas, i.e. the software conducts a separate standard random-effects meta-analysis for each dataset. As in the “metafor” package,⁵ MetaNSUE estimates τ^2 using the REML for its statistical advantages.^{2,10} Remember, however, that we had to use MLE rather than REML in the imputations, due to the reasons discussed above. The covariates included in the MLE step are also included here to allow testing meta-regressions or other linear models.

2.5 Rubin's rules

After completing a meta-analysis for each imputation dataset, MetaNSUE combines the resulting statistics using Rubin's rules⁶

$$\begin{aligned} \beta_{\text{combined}} &= \text{mean}(\beta) \\ \sigma_{\text{combined}}^2 &= \text{mean}(\sigma^2) + \left(1 + \frac{1}{n_{\text{imp}}}\right) \cdot \text{var}(\beta^2) \end{aligned} \quad (11)$$

Note that in case the model has more than one coefficient, the formula to combine the variance-covariance matrices can be derived from the formula to combine the variance of a linear hypothesis (specified with the hypothesis matrix or vector C) as follows

$$\begin{aligned} \sigma_{C,\text{combined}}^2 &= \text{mean}(\sigma_C^2) + \left(1 + \frac{1}{n_{\text{imp}}}\right) \cdot \text{var}(\beta_C) \\ &= C \times \left[\text{mean}(\Lambda) + \left(1 + \frac{1}{n_{\text{imp}}}\right) \cdot \text{cov}(\beta) \right] \times C^T \end{aligned} \quad (12)$$

and thus

$$\Lambda_{\text{combined}} = \text{mean}(\Lambda) + \left(1 + \frac{1}{n_{\text{imp}}}\right) \cdot \text{cov}(\beta) \quad (13)$$

2.6 Hypothesis testing

Then, MetaNSUE tests the hypothesis specified by the meta-analysts as follows

$$\begin{aligned} \beta_{C,\text{combined}} &= C \times \beta_{\text{combined}} \\ \sigma_{C,\text{combined}}^2 &= C \times \Lambda_{\text{combined}} \times C^T \\ Z_{C,\text{combined}} &= \frac{\beta_{C,\text{combined}}}{\sigma_{C,\text{combined}}} \end{aligned} \quad (14)$$

Finally, the software combines the heterogeneity τ and I statistics using the mean, estimates the combined H^2 from the combined I^2 and combines Cochran's Q statistic using the formula for χ^2 statistics

$$F_{Q,\text{combined}} = \frac{\frac{\text{mean}(Q)}{df} - \frac{n_{\text{imp}}+1}{n_{\text{imp}}-1} \cdot r}{1+r}, \quad \text{with} \begin{cases} df_1 = df \\ df_2 = \frac{n_{\text{imp}}-1}{df^{n_{\text{imp}}}} \cdot \left(1 + \frac{1}{r}\right)^2 \end{cases} \quad (15)$$

where df are the degrees of freedom of the model, and

$$r = \left(1 + \frac{1}{n_{\text{imp}}}\right) \cdot \text{var}(\sqrt{Q}) \quad (16)$$

Note that Cochran's Q follows a χ^2 distribution with df degrees of freedom, but its combined statistic follows an F distribution with df_1 and df_2 degrees of freedom. However, for "visual" compatibility with standard meta-analyses, the software converts the F statistic into a χ^2 statistic with df degrees of freedom

As noted earlier, if the MetaNSUE meta-analyst wants to test another hypothesis, he/she must repeat the process from the MLE step.

3 Validation

For the preliminary version of MetaNSUE, we conducted a small simulation work that showed that the estimation of the effect size was unbiased or, at least, substantially less biased than when discarding the studies with NSUEs or when considering that NSUEs were null.³

Here, we report a substantially more exhaustive simulation work. We have examined the bias, the root mean squared error (RMSE) and the false-positive rate, for both the effect size and the heterogeneity, under four scenarios (null vs. medium effect size, no heterogeneity vs. moderate heterogeneity), for MetaNSUE and three alternative strategies.

Each simulated meta-analysis included 40 one-sample studies with random sample sizes uniformly distributed between 20 and 40. To simulate studies with null effect size and no heterogeneity, we generated the t -values of the studies according to a t -distribution. To simulate studies with null effect size but moderate heterogeneity, we generated the t -values according to a non-central t -distribution with random non-centrality parameters normally distributed around 0 with 1.5^2 variance. To simulate studies with medium effect size and no heterogeneity, we generated the t -values according to a non-central t -distribution with non-centrality parameter = 3. Finally, to simulate studies with medium effect size and moderate heterogeneity, we generated the t -values according to a non-central t -distribution with random non-centrality parameters normally distributed around 3 with 1.5^2 variance.

We first conducted the meta-analyses using all data, and saved their estimations of effect size (Hedges' g) and heterogeneity (I^2) as the ground truth. Afterwards, we removed the t -values of non-statistically significant studies, miming NSUEs, and re-analyzed the data with several strategies. These included MetaNSUE, a MetaNSUE-like strategy in which the imputations were first combined using Rubin's rules and afterwards meta-analyzed (rather than the other way round), the discarding of studies with NSUEs, and the consideration that NSUEs are null.

We calculated the bias of a strategy as the mean difference between their estimations and the estimations obtained in the meta-analyses of all data, and the RMSE as the square root of the mean squared difference. Note that the latter is the square root of the sum of the squared bias and the variance of the estimations. We included 15,000 meta-analyses of studies with null effect size and no heterogeneity, 15,000 meta-analyses of studies with null effect size but moderate heterogeneity, 15,000 meta-analyses of studies with medium effect size and no heterogeneity, and 15,000 meta-analyses of studies with medium effect size and moderate heterogeneity. We had previously removed meta-analyses without studies with NSUEs (because the results would be the same with any strategy) and meta-analyses with only studies with NSUEs (because they would not be conducted).

Finally, we calculated the observed p -values of the test of the main hypothesis (i.e. whether the effect size is not null) and the test of heterogeneity. For the test of the main hypothesis, we included 60,000 meta-analyses of studies with null effect-size (30,000 without heterogeneity and 30,000 with moderate heterogeneity), and we counted how many of them had a z -value higher than the z -value associated to $p = 0.05, 0.01, 0.005$ and 0.001 (two-tailed). For the test of heterogeneity, we included 60,000 meta-analyses of studies with no heterogeneity (30,000 with null effect size and 30,000 with medium effect size), and we counted which proportion of them had a Q statistic higher than the χ^2 statistic associated to $p = 0.05, 0.01, 0.005$, and 0.001 . These percentages of meta-analyses were the observed p -values, and we compared them with the expected p -values (0.05, 0.01, 0.005, and 0.001). For this analysis, we did

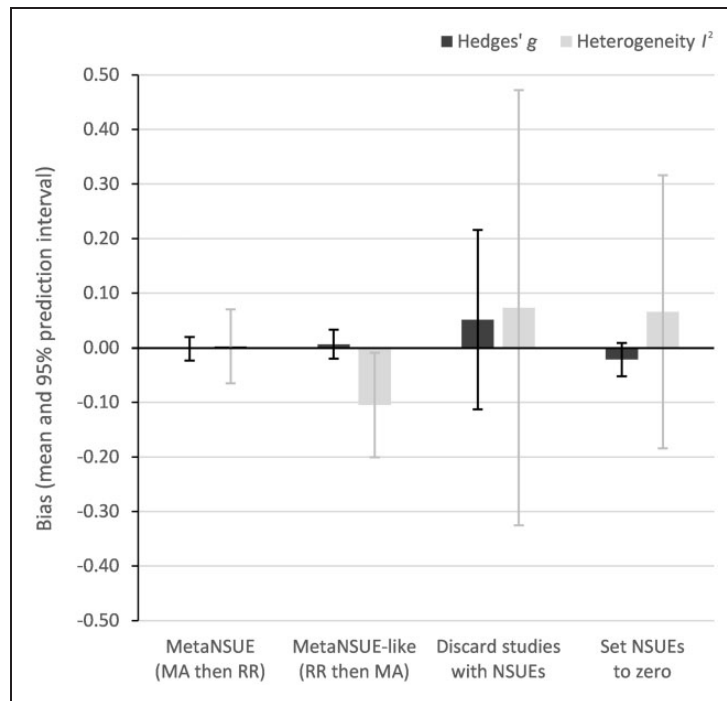


Figure 3. Bias (mean and 95% prediction interval) in the estimation of the main hypothesis Hedges' g and the heterogeneity statistic I^2 in the different approaches for studies with non-statistically significant unreported effects (NSUEs). MA: meta-analysis; RR: Rubin's rules.

not remove meta-analyses without studies with NSUEs or meta-analyses with only studies with NSUEs; z - and Q -values for studies with only NSUEs in the "Discard NSUEs" approach were imputed as null.

3.1 Results

As shown in Figure 3 and Table 2, the bias in the estimation of the effect size or the heterogeneity was null or negligible with MetaNSUE. The MetaNSUE-like strategy underestimated the heterogeneity, and to a much lesser extent, inflated the effect size. Discarding the studies with NSUEs inflated the effect size, and it inflated the heterogeneity when the effect size was null while it shrank it when the effect size was medium. The outcomes were opposite when we considered that NSUEs were zero: this strategy shrank the effect size, and it shrank the heterogeneity when the effect size was null, while it inflated it when the effect size was medium. In addition, the two latter strategies were substantially imprecise (i.e. large RMSE and prediction intervals of the bias).

The observed p -values of the test of the main hypothesis (Table 3) were globally accurate for MetaNSUE, although moderately conservative in the absence of heterogeneity (observed p -values three times smaller than expected) and moderately liberal in the presence of moderate heterogeneity (observed p -values three times larger than expected). The MetaNSUE-like strategy yielded liberal p -values, especially in the presence of moderate heterogeneity (observed p -values eight times larger than expected). Finally, the observed p -values were strongly liberal when discarding studies with NSUEs and strongly conservative when considering that NSUEs are null, especially in the absence of heterogeneity.

The observed p -values of the test of heterogeneity were strongly conservative in all cases, except when the effect size was medium and we considered NSUEs null, in which case they were strongly liberal (Table 3).

4 Worked example (continuation)

Resuming the worked example presented in the introduction of the paper, we should first create an object of class "nsue" using the R function "smc_from_t" (standardized mean change from t -value), and then we should simply conduct the meta-analysis with the R function "meta":

```
R > t <- c(3.4, NA, NA, NA, NA,
```

Table 2. Bias and root mean square error (RMSE) in the estimation of the main hypothesis Hedges' g and the heterogeneity statistic I^2 in the different approaches for studies with non-statistically significant unreported effects (NSUEs), separately for studies with null or medium effect size and no or moderate heterogeneity.

			MetaNSUE (MA then Rubin's rules)	MetaNSUE-like (Rubin's rules then MA)	Discard studies with NSUEs	Set NSUEs to zero
Estimation of g						
$g = 0$	$I^2 = 0.07^a$	Bias	No	No	No*	No
		RMSE	0.03	0.03	0.27	0.02
$g = 0.53$	$I^2 = 0.64$	Bias	No	No	No*	No
		RMSE	0.03	0.04	0.14	0.03
	$I^2 = 0.06^a$	Bias	No	+0.01	+0.06*	-0.04
		RMSE	0.01	0.01	0.06	0.04
$I^2 = 0.61$	Bias	No	+0.02	+0.15*	-0.05	
	RMSE	0.01	0.02	0.15	0.05	
Estimation of I^2						
$g = 0$	$I^2 = 0.07^a$	Bias	No	-0.07*	+0.51*	-0.07*
		RMSE	0.07	0.11	0.64	0.12
$g = 0.53$	$I^2 = 0.64$	Bias	-0.01	-0.17*	+0.23*	-0.19*
		RMSE	0.04	0.19	0.25	0.23
	$I^2 = 0.06^a$	Bias	+0.02	-0.03	-0.06*	+0.41*
		RMSE	0.08	0.07	0.11	0.43
$I^2 = 0.61$	Bias	-0.01	-0.14*	-0.38*	+0.11*	
	RMSE	0.07	0.17	0.42	0.13	

Note: We marked with an asterisk biases ≥ 0.05 or RMSE ≥ 0.10 . Data are based on 60,000 simulated meta-analyses of 40 one-sample studies with random sample sizes uniformly distributed between 20 and 40. The t -values of the studies followed a t -distribution in 15,000 meta-analyses, a non-central t -distribution with non-centrality parameter $N(0, 1.5^2)$ in 15,000 meta-analyses, a non-central t -distribution with non-centrality parameter = 3 in 15,000 meta-analyses, and a non-central t -distribution with non-centrality parameter $N(3, 1.5^2)$ in the remaining 15,000 meta-analyses. These simulations do not include meta-analyses with only NSUEs or meta-analyses with no NSUEs. MA: meta-analysis.

^aWe simulated these meta-analyses with no heterogeneity. By chance, in some meta-analyses, the variability between studies was exactly that expected from sampling error ($I^2 = 0$), but in some it was smaller ($I^2 < 0$) and in others it was larger ($I^2 > 0$). However, negative I^2 statistics are usually truncated to 0, and thus mean I^2 across meta-analyses was slightly larger than 0.

```
2.8, 2.1, 3.1, 2.0, 3.4)
R > n <- c(40, 20, 22, 24, 18,
30, 25, 30, 16, 22)
R > meta(smc_from_t(t, n))
```

Note that we have indicated NSUEs with NA.

To enrich the picture, we could easily add a forest plot with the R function "forest" (Figure 4), a jack-knife analysis with the function "leave1out," meta-regressions by moderators of interest, and a meta-regression by standard errors and/or a funnel plot with the R functions "metabias" and "funnel" (Figure 5) to detect potential reporting bias:

```
R > leave1out(smc_from_t(t, n))
R > m <- meta(smc_from_t(t, n))
R > forest(m)
R > funnel(m)
```

Please see the Supplementary Material for details of the graphical steps to obtain these analyses using the GUI.

Using the example data provided in the Supplementary Material, MetaNSUE would estimate a statistically significant standardized mean change (effect size = 0.42, 95% CI: 0.27 – 0.57; $z = 5.6$, $p < 0.001$), with low between-study heterogeneity ($I^2 = 3\%$) and low risk of reporting/publication bias ($Z = -0.4$, $p = 0.676$).

Table 3. Expected and observed p -values in the tests of the main hypothesis and heterogeneity in the different approaches for studies with non-statistically significant unreported effects (NSUEs), separately for studies with no or moderate heterogeneity, and with null or medium effect size.

			Expected (two-tailed)	MetaNSUE (MA then Rubin's rules)	MetaNSUE-like (Rubin's rules then MA)	Discard studies with NSUEs	Set NSUEs to zero
Test of the main hypothesis (z)							
$g = 0$	$I^2 = 0.05^a$	$z = 1.96$	0.050	0.014	0.041	0.201	<0.001*
		$z = 2.58$	0.010	0.002	0.014	0.199*	<0.001*
		$z = 2.81$	0.005	0.002	0.010	0.192*	<0.001*
		$z = 3.29$	0.001	<0.001	0.003	0.104*	<0.001*
	$I^2 = 0.64$	$z = 1.96$	0.050	0.071	0.132	0.081	0.052
		$z = 2.58$	0.010	0.026	0.063*	0.035	0.010
		$z = 2.81$	0.005	0.017	0.048*	0.026*	0.005
		$z = 3.29$	0.001	0.007*	0.026*	0.016*	0.001
Test of heterogeneity (Q)							
$g = 0$	$I^2 = 0.05^a$	$Q = 54.6$	0.050	0.004*	<0.001*	<0.001*	<0.001*
		$Q = 62.4$	0.010	0.001*	<0.001*	<0.001*	<0.001*
		$Q = 65.5$	0.005	<0.001*	<0.001*	<0.001*	<0.001*
		$Q = 72.1$	0.001	<0.001*	<0.001*	<0.001*	<0.001*
$g = 0.53$	$I^2 = 0.06^a$	$Q = 54.6$	0.050	0.001*	0.023	<0.001*	0.925*
		$Q = 62.4$	0.010	<0.001*	0.004	<0.001*	0.830*
		$Q = 65.5$	0.005	<0.001*	0.002	<0.001*	0.777*
		$Q = 72.1$	0.001	<0.001*	<0.001	<0.001*	0.635*

Note: We marked with an asterisk those observed p -values that were five or more times larger or smaller than the expected p -values. Example of interpretation: under the main null hypothesis and in the absence of heterogeneity, a z -value of 1.96 should correspond to a two-tailed p -value = 0.05, but it erroneously corresponds to p -value = 0.20 (i.e. non-statistically significant) if studies with NSUEs are discarded, and to p -value < 0.001 (i.e. very statistically significant) if studies with NSUEs are assumed to have a null effect size. Data are based on 90,000 simulated meta-analyses of 40 one-sample studies with random sample sizes uniformly distributed between 20 and 40. The t -values of the studies followed a t -distribution in 30,000 meta-analyses, a non-central t -distribution with non-centrality parameter $\sim N(0, 1.5^2)$ in 30,000 meta-analyses, and a non-central t -distribution with non-centrality parameter = 3 in the remaining 30,000 meta-analyses. z - and Q -values for studies with only NSUEs in the "Discard NSUEs" approach were imputed as null. MA: meta-analysis.

^aWe simulated these meta-analyses with no heterogeneity. By chance, in some meta-analyses, the variability between studies was exactly that expect from sampling error ($I^2 = 0$), but in some it was smaller ($I^2 < 0$) and in others it was larger ($I^2 > 0$). However, negative I^2 statistics are usually truncated to 0, and thus mean I^2 across meta-analyses was slightly larger than 0.

5 Discussion

This paper presents a novel meta-analytic method that can include studies that report that the outcome was not statistically significant but do not report the actual outcome. With this method, we hope that both R and non-R meta-analysts are able to conduct robust MetaNSUE meta-analyses rather than, incorrectly, excluding NSUE studies or assuming that they have a null effect size.

In the preliminary version of the method,³ we had already found that the method yields unbiased estimations of the effect size. Now we also show that the p -values of the test of the main hypothesis are (globally) accurate and that the estimation of the heterogeneity is (nearly) unbiased. Conversely, we show that the simpler approaches have strong biases in the estimation of both the effect size and the heterogeneity and poorly control the false-positive rate of the test of the main hypothesis.

We would like to clarify that the simpler approaches did not bias the effect size of meta-analyses with null effect size, but we think that this may be simply because in this situation, biases increasing the effect size and biases decreasing the effect size counteract. For instance, if a strategy tends to inflate the effect size, it will increase positive effect sizes but also decrease negative effect sizes, and if a strategy tends to shrink the effect size, it will decrease positive effect sizes but also increase negative effect sizes.

However, discarding studies with NSUEs led to a severe inflation of the heterogeneity and to strongly liberal p -values in the test of the main hypothesis. We expected this behavior, as the effect sizes were only either very negative or very positive (we had discarded studies with null to medium effect sizes). Oppositely, considering that NSUEs are null led to a severe shrinkage of the heterogeneity and to strongly conservative p -values in the test of

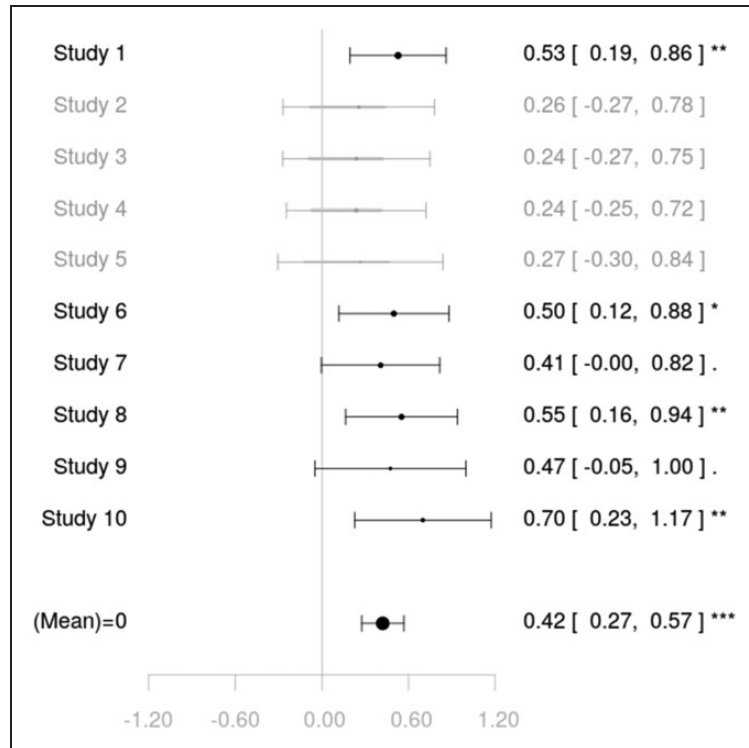


Figure 4. Forest plot of the example meta-analysis. MetaNSUE plots the imputations of the effects of studies with non-statistically significant unreported effects (NSUEs) in gray, and marks with asterisks statistically significant results.

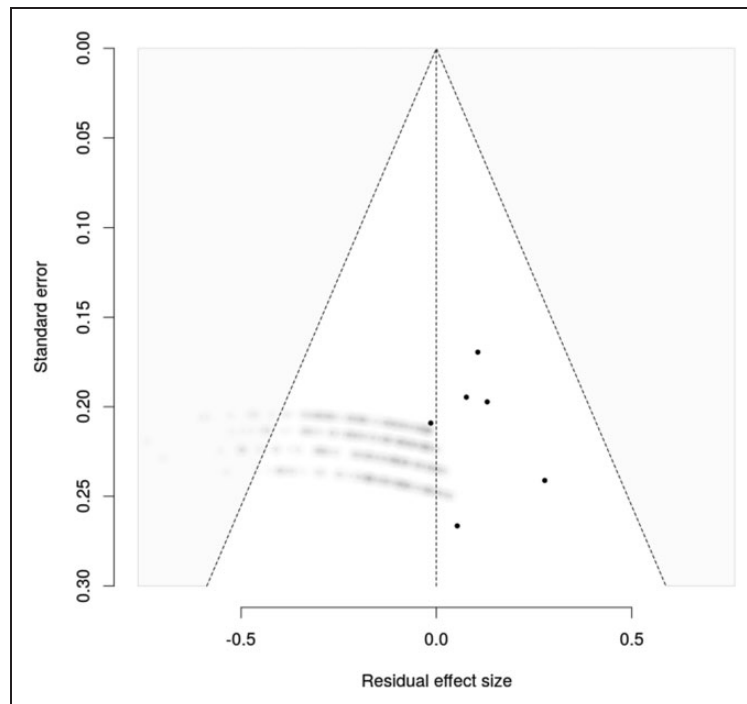


Figure 5. Funnel plot of the example meta-analysis. MetaNSUE plots the imputations of the effects of studies with non-statistically significant unreported effects (NSUEs) in gray.

the main hypothesis. We might explain this behavior by the fact that a portion of the studies had the same, null effect size (we had considered that NSUEs were zero). The presence of moderate heterogeneity reduced the lack of control of the false-positive rate; we assume that this is at least partly due to a smaller number of studies with NSUEs.

Finally, the pattern of biases in the estimation of the heterogeneity was opposite in meta-analyses with medium effect size: discarding studies with NSUEs shrank the estimation of heterogeneity, while considering that NSUEs are zero inflated it. We again expected this behavior, because in the first case, effect sizes were only very positive (we had removed studies with positive but small to medium effect sizes), and in the second case, the effect sizes were either very positive or null.

In the validation, we also tried a modified version of MetaNSUE, which in the text we referred to as MetaNSUE-like. In this modified version, we first combined the multiple imputations using Rubin's rules, separately for each study with NSUEs, and then conducted a standard meta-analysis. We must highlight that a previous study¹¹ found this order more accurate for imputing missing individual data on covariates. Our results were opposite, but the scenarios of that study and ours are completely different. Specifically, we found that this strategy tended to underestimate the heterogeneity, to make p -values of the test of the main hypothesis liberal and, to a much lesser extent, to inflate the effect size. We think that the origin of these biases might be the fact that combining the multiple imputations using Rubin's rules at the study level "falsely" increases the within-study variance of the studies with NSUEs. If the global within-study variance is larger, the meta-analysis software assumes that the variability between studies is more due to sampling error and fewer related to between-study heterogeneity, thus underestimating the latter, which in turn might explain the liberal p -values. In addition, the strategy only increases the within-study variance of the studies with NSUEs, decreasing their weight, and thus partially adopting the strategy of discarding studies with NSUEs.

MetaNSUE has a number of limitations. First, the test of heterogeneity is strongly conservative. This means that MetaNSUE meta-analysts will hardly be wrong when stating that there is heterogeneity in a meta-analysis (i.e. it has a very low false positive rate), but it also means that they will fail to detect heterogeneity in many meta-analyses. For this reason, we strongly suggest that MetaNSUE meta-analysts focus more on the I^2 statistic than on the Q p -value to assess heterogeneity. We indeed even include this recommendation in the output of the software. Second, the p -values of the main hypothesis were moderately conservative in the absence of heterogeneity and moderately liberal in the presence of moderate heterogeneity. We hypothesize that this behaviour may be due to an imperfect estimation of the heterogeneity. Nevertheless, the p -values were globally accurate, i.e. the false-positive rate was that expected for the significance level. Third, the method has room for improvement in several steps. For example, an improvement could be an unbiased estimation of the heterogeneity in a single step, rather than using a combination of MLE and multiple imputations. Fourth, MetaNSUE attempts to solve the lack of reporting of statistics when effects are not statistically significant, but not the complete lack of reporting (e.g. as in the file drawer problem). However, we have included standard tools to assess reporting bias, such as funnel plots and tests. Finally, MetaNSUE software currently assumes a truncated normal distribution and has only built-in functions for standardized mean changes, standardized mean differences and correlations. However, R meta-analysts may easily adapt the software to other distributions and measures of effect size, because the software allows the definition of personalized variance, imputation and back-transforming functions. We are completely open to increase further the flexibility of the software if meta-analysts need to use other personalized functions, and indeed, we will probably add built-in functions for other measures of effect size in the future.

Acknowledgements

We would like to thank Dr. Roby Joehanes from Harvard Medical School for providing valuable help in some of the numerical improvements explained in the text.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by PFIS Predoctoral Contract FI16/00311, Miguel Servet Research Contract MS14/00041, and Research Project PI14/00292 from the Plan Nacional de I + D + i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund (FEDER). The funders

had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

ORCID iD

Joaquim Radua  <http://orcid.org/0000-0003-1240-5438>

Supplemental material

Supplemental material for this article is available online.

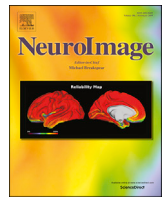
References

1. Radua J, Mataix-Cols D, Phillips ML, et al. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *Eur Psychiatry* 2012; **27**(8): 605–611.
2. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat* 2005; **30**: 261–293.
3. Radua J, Schmidt A, Borgwardt S, et al. Ventral striatal activation during reward processing in psychosis: a neurofunctional meta-analysis. *JAMA Psychiatry* 2015; **72**: 1243–1251.
4. Rubin D. *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons, 1987.
5. Viechtbauer W. Conducting meta-analyses in r with the metafor package. *J Stat Softw* 2010; **36**: 1–48.
6. Li K, Meng X, Raghunathan T, et al. Significance levels from repeated p-values with multiply-imputed data. *Stat Sin* 1991; **1**: 65–92.
7. Fox J. *Applied regression analysis and generalized linear models*, 2nd ed. Thousand Oaks: Sage, 2008.
8. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**: 24–36.
9. Schnedler W. Likelihood estimation for censored random vectors. *Econ Rev* 2005; **24**: 195–217.
10. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Meth* 2016; **7**: 55–79.
11. Burgess S, White IR, Resche-Rigon M, et al. Combining multiple imputation and meta-analysis with individual participant data. *Stat Med* 2013; **32**: 4499–4514.



Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM

Anton Albajes-Eizagirre^{a,b}, Aleix Solanes^{a,b,c}, Eduard Vieta^{b,c,d,e}, Joaquim Radua^{a,b,c,f,g,*}

^a FIDMAG Germanes Hospitalàries, Sant Boi de Llobregat, Barcelona, Spain

^b Mental Health Research Networking Center (CIBERSAM), Madrid, Spain

^c Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

^d Universitat de Barcelona, Barcelona, Spain

^e Clinical Institute of Neuroscience, Hospital Clínic de Barcelona, Barcelona, Spain

^f Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

^g Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

ARTICLE INFO

Keywords:

Coordinate-based meta-analysis
Tests for spatial convergence
Familywise error rate
Activation likelihood estimation
Seed-based d mapping
Signed differential mapping

ABSTRACT

Coordinate-based meta-analyses (CBMA) are very useful for summarizing the large number of voxel-based neuroimaging studies of normal brain functions and brain abnormalities in neuropsychiatric disorders. However, current CBMA methods do not conduct common voxelwise tests, but rather a test of convergence, which relies on some spatial assumptions that data may seldom meet, and has lower statistical power when there are multiple effects. Here we present a new algorithm that can use standard voxelwise tests and, importantly, conducts a standard permutation of subject images (PSI). Its main steps are: a) multiple imputation of study images; b) imputation of subject images; and c) subject-based permutation test to control the familywise error rate (FWER). The PSI algorithm is general and we believe that developers might implement it for several CBMA methods. We present here an implementation of PSI for seed-based d mapping (SDM) method, which additionally benefits from the use of effect sizes, random-effects models, Freedman-Lane-based permutations and threshold-free cluster enhancement (TFCE) statistics, among others. Finally, we also provide an empirical validation of the control of the FWER in SDM-PSI, which showed that it might be too conservative. We hope that the neuroimaging meta-analytic community will welcome this new algorithm and method.

1. Introduction

Meta-analyses are essential to summarize the wealth of findings from voxel-based neuroimaging studies, as well as to assess potential reporting bias, between-study heterogeneity or the influence of moderators (Radua and Mataix-Cols, 2012). However, meta-analytic researchers in voxel-based neuroimaging cannot apply standard statistical procedures without having the three-dimensional (3D) statistical images of the results of the studies, which are unfortunately unavailable for most studies. For instance, in a recent meta-analysis, the 3D statistical images were available in only nine out of the 50 studies, i.e., 41 of the studies only reported the coordinates and t-values of the peaks of statistical significance (Wise et al., 2016). To overcome this problem, the neuroimaging community developed alternative procedures that only require the coordinates of the peaks of the clusters of statistical significance (Radua and

Mataix-Cols, 2009; Radua et al., 2012, 2014; Turkeltaub et al., 2002, 2012; Laird et al., 2005; Eickhoff et al., 2009, 2012; Wager et al., 2007; Costafreda et al., 2009; Costafreda, 2012; Kang et al., 2011, 2014; Yue et al., 2012; Montagna et al., 2017). Many meta-analysts have called these methods coordinate-based meta-analyses (CBMA) (Radua and Mataix-Cols, 2012).

An important feature of CBMA is the use of a statistical procedure that, instead of testing whether the effects are not null, tests whether the reported findings tend to converge in some brain regions (Albajes-Eizagirre and Radua, 2018). Unfortunately, we have recently showed that the test for convergence used by CBMA might have two drawbacks. First, it relies on several spatial assumptions but data may seldom meet them, leading to either conservative or liberal results. Second, its statistical power decreases when there are multiple findings (Albajes-Eizagirre and Radua, 2018).

* Corresponding author. Division of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, PO 69. 16 De Crespigny Park, London, SE5 8AF, UK.

E-mail address: Joaquim.Radua@kcl.ac.uk (J. Radua).

<https://doi.org/10.1016/j.neuroimage.2018.10.077>

Received 21 August 2018; Received in revised form 10 October 2018; Accepted 29 October 2018

Available online 30 October 2018

1053-8119/© 2018 Elsevier Inc. All rights reserved.

To overcome these drawbacks, we developed a new CBMA algorithm that can use standard univariate voxelwise tests. In other words, it can test whether effects are not null in a given voxel, rather than whether findings tend to converge around the voxel. We must note at this point that there are two standard testing approaches in voxel-based neuroimaging: parametric tests, and permutation tests, but a recent study showed that the former might be conservative for voxel-based statistics and invalid for cluster-based statistics, whereas the latter correctly controls the FWER (Eklund et al., 2016). We aimed to develop a correct test and thus chose the permutation of subject images. For this reason, we then call the new algorithm “Permutation of Subject Images” (PSI) CBMA. We acknowledge that a sign-flipping permutation of study images would be quicker than a subject-based permutation and could similarly test whether effects are not null. However, the improvement in computation time would be small while there would be a decrease in the accuracy of the estimation of p-values (we expand this subject in the Discussion).

The algorithm is general and we believe that developers could implement it to several current CBMA methods such as Activation Likelihood Estimation (ALE) (Turkeltaub et al., 2002, 2012; Laird et al., 2005; Eickhoff et al., 2009, 2012) or Multilevel Kernel Density Analysis (MKDA) (Wager et al., 2007). Here we present its implementation for Anisotropic Effect-Size Seed-based d Mapping (AES-SDM) (Radua et al., 2012, 2014) for its key advantages, e.g., it imputes a 3D effect-size image of each study and then fits standard meta-analytic random-effects models. In the context of CBMA, the use of effect-sizes and random-effects models were associated with increased reliability and performance in a recent methodological study (Bossier et al., 2017). In addition, AES-SDM accounts for both increases and decreases of the measure (e.g., activations and deactivations) so that contradictory findings cancel each other (Radua and Mataix-Cols, 2009), it considers the irregular local spatial covariance of the different brain tissues (Radua et al., 2014), and allows the simultaneous inclusion of peak coordinates and available 3D statistical images, substantially increasing the statistical power (Radua et al., 2012). The major change of the new SDM-PSI method is the imputation of subject images to allow a subject-based permutation test, in an identical fashion to that of FSL “randomize” tool (Winkler et al., 2014) or SPM Statistical NonParametric Mapping toolbox (Nichols and Holmes, 2002). Thus, SDM-PSI, FSL or SPM test whether the activation of a voxel is different from zero, while standard CBMA test whether studies report activations in the voxel more often than in other voxels. Other improvements are a less biased estimation of the population effect size, the possibility of using threshold-free cluster enhancement (TFCE) statistics (Smith and Nichols, 2009), and the multiple imputation of study images, avoiding the biases associated with single imputation (Rubin, 1987).

We present the novel algorithm and method in two successive sections of the manuscript. First, we describe the general PSI algorithm beyond SDM, and second, we detail the specific implementation of PSI for SDM. With this division, we aim to both make the manuscript easier to read, and to highlight the fact that other developers could indeed implement PSI to CBMA methods other than SDM. In a third section of the manuscript, we report the empirical validations of SDM-PSI. We hope that the neuroimaging meta-analytic community will welcome this new algorithm and method.

2. The PSI algorithm

2.1. Overview

The main pillar of the PSI algorithm is to conduct a permutation test of the subject images, in an identical fashion to that of FSL “randomize” tool (Winkler et al., 2014) or SPM Statistical NonParametric Mapping toolbox (Nichols and Holmes, 2002). Of course, it is impossible to recreate the original subject images of the included studies, neither from the peak information reported in the papers nor from the 3D statistical study images. However, we show later that there is no need to recreate

the exact original subject images. If the imputation algorithm meets some conditions, the subject-based permutation test will be correct even if the similarities with the original subject images are scarce.

The main steps of the PSI method, which we extend below, are:

1. For each study from which only peak information is available, impute several study images that show realistic local spatial covariance and that adequately cover the different possibilities within the uncertainty. Of course, the algorithm does not need to impute images for the studies from which images are available. We name “imputed dataset” each set of images, one per study.
2. For each study, impute subject images that show realistic local spatial covariance. Then adapt them to the different imputed study images, so that the group analysis of the subject images of an imputed dataset returns the study images of that imputed dataset. For example, for normally distributed data, impute subject images once for all imputations, and then scale them to the different imputed study images.
3. Perform a subject-based permutation test as follows:
 - a. Create one random permutation of the subjects and apply it to the subject images of the different imputed datasets.
 - b. Separately for each imputed dataset, conduct a group analysis of the permuted subject images to obtain one study image per study, and then conduct a meta-analysis of the study images to obtain one meta-analysis image.
 - c. Use Rubin’s rules to combine the meta-analysis images from the different imputed datasets to obtain a combined meta-analysis image (Li et al., 1991).
 - d. Save a maximum statistic from the combined meta-analysis image (e.g., the largest z-value).
 - e. Go to step a).
 - f. After enough iterations of steps a) to d), use the distribution of the maximum statistic to threshold the combined meta-analysis image obtained from unpermuted data.

Thus, as in FSL or SPM (Winkler et al., 2014; Nichols and Holmes, 2002), an iteration of the permutation test consists in repeating the analysis using the permuted subject images and saving a maximum statistic from the images derived from the permuted images. See Fig. 1 for a simplified flow of the algorithm.

2.2. Imputation of study images

We define a “study image” as the 3D statistical image of the contrast of interest in the group-level analysis conducted in the study. For example, SPM and FSL study images have t-values and their names are similar to “spmT_0001.nii” or “design_tstat1.nii.gz”. CBMA methods do

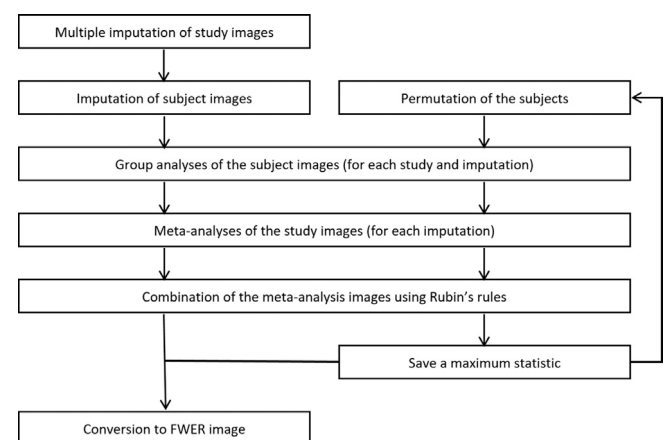


Fig. 1. Simplified flow of the PSI algorithm. Footnote: FWER: familywise error rate.

not use these raw study images, but transformations or imputations thereof. For example, AES-SDM uses images of effect sizes (Radua et al., 2012), MKDA uses binary images representing regions close to peaks (Wager et al., 2007), and ALE uses images of the likelihood that peaks lie around each voxel (Turkeltaub et al., 2002).

Transformation of raw study images into the study images used by a CBMA may be associated with some error related to numerical precision and spatial interpolation, but this should be negligible and we can safely ignore it. Conversely, imputation of study images from the scarce information reported as peak coordinates in the papers is associated with a substantial amount of uncertainty.

For example, when the raw study image of t-values is available, AES-SDM software can convert it straightforwardly and safely into an image of effect-sizes with negligible error. Similarly, if we took the liberty to redefine MKDA “closeness” as “belonging to the cluster of the peak”, MKDA software could straightforward use the binary image that indicates which voxels are statistically significant. Conversely, when only peak information is available, AES-SDM software must conduct a progressive estimation of the effect-size of the voxels close to the reported peaks, which inevitably introduces a non-negligible amount of uncertainty. Similarly, MKDA only relies on the distance between a voxel and the peak, ignoring the real shape of the cluster.

One novelty of PSI consists of imputing the study images several times, adequately covering this uncertainty and avoiding the biases associated with single imputation (Rubin, 1987). For SDM-PSI, this means imputing several effect sizes for each voxel, covering the different effect sizes that a voxel could have had in the (unavailable) raw study image. For MKDA-PSI, it could mean impute many times whether a voxel was part of the cluster or not, and the more likely a voxel is to have been part of the cluster in the raw study image, the more times MKDA-PSI would impute it as part of the cluster.

Importantly, the values imputed for a voxel must follow a statistical distribution in accordance with the known information and its uncertainty. For SDM-PSI, the mean and standard deviation of the effect-sizes imputed for a voxel must match the estimated effect-size of the voxel and its standard error, and the effect-sizes cannot be statistically significant in non-statistically significant voxels. For MKDA-PSI, the proportion of imputations in which the voxel is part of the cluster could match the probability that the voxel is part of the cluster. Otherwise, the imputations would not be in accordance with the known information.

In addition, the voxels must show a realistic local spatial structure to avoid a distortion of the clustering of statistically significant voxels, which would invalidate not only cluster-based statistics, but also voxel-based statistics as far as a cluster extent threshold is applied. We acknowledge that the word “realistic” is ambiguous, but we show in the validations that simply forcing some positive correlation between adjacent voxels may be enough to control the FWER, with only a few exceptions when using cluster-based statistics (Table 1).

2.3. Imputation of subject images

As stated earlier, it is not possible to recreate exactly the original subject images. However, as we show in Fig. 2 and the Supplement, this does not seem to prevent a correct permutation test as long as the values of a study in a voxel show a perfect correlation between any two imputed datasets. For example, in a normally distributed voxel, the Pearson correlation between the subject values of a study in a given imputed dataset and the subject values of the same study in another imputed dataset must be one. Otherwise, there would be an inflation of the variance between imputations, and this would lead to erroneous increases of the statistical significance.

In addition, the voxels must show a realistic local spatial structure to ensure that the study images, obtained from the group analysis of the permuted subject images, have a local spatial structure as similar as possible to the unpermuted study images, for the same reasons described above for study images.

2.4. Permutations

Following standard procedures (Winkler et al., 2014), PSI methods must randomly assign “+1” or “-1” to each subject of a one-sample study, or randomly reassign each of the subjects of a two-sample study to one of the two groups. With these permutations, we remove the potential effects present in the unpermuted images, and thus the meta-analysis images resulting from permuted subject images represent the outcome of many simulated meta-analyses of studies with no effects. For example, one-sample meta-analysis tests whether the value of a voxel is truly different from zero. Under the null hypothesis, we assume that the value of the voxel is zero in the population, and that the value in our data is different from zero only due to chance, and thus, it is as likely to be greater than zero as to be lower than zero. This is the reason why PSI methods must randomly multiply the value of the voxel by “+1” or “-1”. Similarly, two-sample studies test whether the value of a voxel is truly different between two groups. Under the null hypothesis, we assume that the value of the voxel is the same in the two groups, and that the value is different between our groups only due to chance. Therefore, subjects could randomly belong to one group or the other. Consequently, PSI methods must randomly re-assign subjects to one of the two groups. For analogous reasons, PSI methods must also swap subjects in a correlation meta-analysis.

Importantly, the permutation must be the same for all imputed datasets. For example, if in an iteration a PSI method assigns a “-1” to subject #3 of study #5, the PSI method will have to multiply the images of subject #3 of study #5 by “-1” in all imputed datasets. As we show in Fig. 2 and the Supplement, if we permuted the subjects differently in each imputed dataset, there would be an inflation of the variance between imputations, and this also would lead to erroneous increases of the statistical significance.

Table 1
Empirical familywise error rate (FWER) as observed in the simulated null meta-analyses.

Statistics			Empirical FWER				
			Global	Small meta-analyses		Large meta-analyses	
				Small studies	All studies	Small studies	All studies
Voxel		1% (0–2%)	0% (0–2%)	1% (0–3%)	0% (0–1%)	0% (0–1%)	
Cluster	z = 2.33	Size	4% (3–5%)	6% (4–9%)	1% (0–2%)	0% (0–1%)	0% (0–1%)
		Mass	11% (9–13%)	19% (15–23%)	2% (1–4%)	1% (0–2%)	0% (0–1%)
	z = 3.09	Size	4% (3–5%)	7% (5–10%)	0% (0–1%)	0% (0–1%)	0% (0–1%)
		Mass	12% (10–14%)	19% (15–23%)	2% (1–4%)	2% (1–4%)	0% (0–2%)
TFCE		5% (4–7%)	8% (6–11%)	1% (0–3%)	1% (0–3%)	0% (0–1%)	

(a) Small studies had 11, 13, 16, 20 or 24 subjects per group; all studies had these sample sizes plus 30, 36, 44, 54 or 67 subjects per group. (b) Small meta-analyses included 10 studies; large meta-analyses included 20 studies. (c) Minimum FWER was always observed in large meta-analyses of all studies; maximum FWER was always observed in small meta-analyses of small studies.

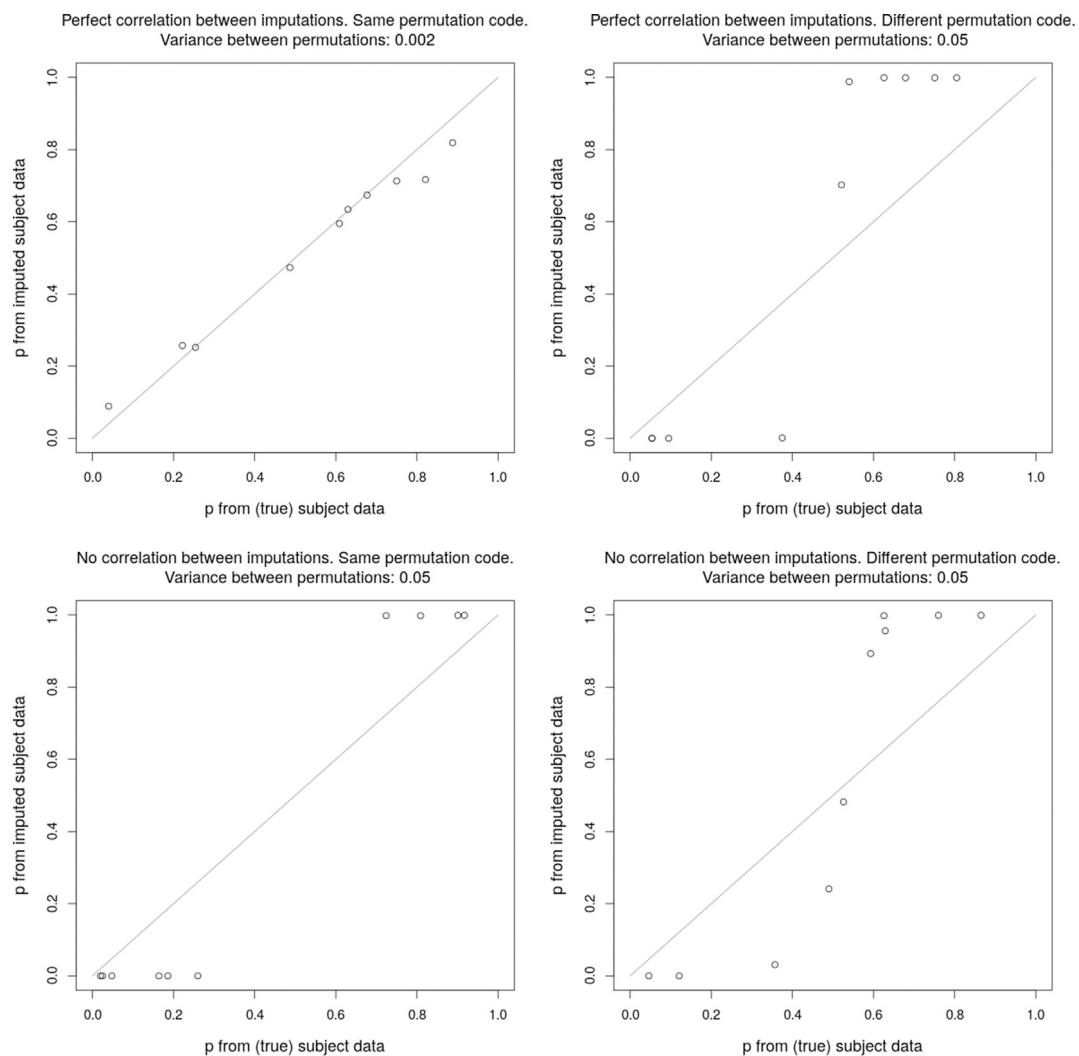


Fig. 2. Effects of the correlation between imputations and the use of variable permutation codes in statistical significance. *Footnote:* Simulation of multiple imputation of effect size, imputation of subject values and permutation test for a single study, forcing a perfect correlation between imputations or not, and using the same or a different permutation code. This figure is the output of the script in R-language provided in the Supplement. Feel free to use that script to check these effects under different parameters.

2.5. Group analysis, meta-analysis and Rubin's rules

In this step, the permuted imputed subject images must be voxelwise combined into study images (one for each study within each imputed dataset), these study images must be voxelwise combined into meta-analysis images (one for each imputed dataset), and these meta-analysis images must be voxelwise combined using Rubin's rules (Li et al., 1991). Group analysis and meta-analysis may vary much depending on the specific CBMA method.

2.6. Maximum statistic test

From each permutation, PSI methods must save a maximum statistic of the combined meta-analysis image, for example the largest value (i.e., the value of the global peak). If we aim a FWER of 5%, we may then consider that a voxel of the unpermuted combined meta-analysis image is statistically significant if it is higher than 95% of these maxima (Nichols and Holmes, 2002; Holmes et al., 1996). Obviously, only 5% of the permuted combined meta-analysis images will have maxima larger than 95% of these maxima, and thus only 5% of the null meta-analyses would erroneously have one or more statistically significant findings. Note that the maximum of the unpermuted combined meta-analysis image (i.e., the first iteration) must be also saved to this null distribution (Phipson and

Smyth, 2010).

2.7. Meta-regression and other linear models

PSI methods can only permute subject images for the main analysis (i.e., the mean). For meta-regression and other linear models, they must conduct the permutation at the study-level, because we only know the value of the moderators at this level. For example, in a meta-regression by the percentage of medicated patients, we may know that 53% patients were medicated in study #1 and 28% patients were medicated in study #2, but we do not know which specific patients were medicated and which were not.

Thus, when conducting a simple meta-regression, PSI methods do not need to impute subject images, permute them and conduct group analysis. Rather, they have to permute the value of the moderator between studies. The reason is that under the null hypothesis, we assume that the value of the voxels is unrelated to the value of the moderator, and that if we observe any relationship between the voxels and the moderator in our data is only due to chance.

The permutation is not as straightforward when there are nuisance variables, and developers can choose among a number of approaches, though a previous comparison of these approaches show that the Freedman-Lane method had optimal statistical properties (Winkler et al., 2014).

3. Implementation of PSI in SDM

3.1. Overview

In this section, we describe how we implemented the PSI method to an existing CBMA method, the AES-SDM (Radua et al., 2012, 2014). We graphically summarize the steps in Figs. 3–5.

We would like to highlight that some of the novelties of the new version of SDM represent an improvement even if the user is not interested in the p-values and thus does not conduct a permutation test. At this regard, the use of maximum-likelihood estimation (MLE) and multiple imputation techniques make the estimation of the population effect sizes substantially less biased than in previous versions of SDM (Radua et al., 2015; Albajes-Eizagirre et al., 2018). The software is freely available at <https://www.sdmproject.com/>.

3.2. Imputation of study images

As in previous versions of SDM, the input data for a study may be either a raw study image or a set of peak coordinates and t-values, and a meta-analysis may combine both types of input (Radua et al., 2012).

Obviously, if the raw study image is available, the software does not need to impute it. The only preprocessing is a conversion of its t-, z- or p-values to effect-sizes and potentially a spatial interpolation to the voxel dimensions and space of the meta-analytic template. Conversion from t-values into Hedge's g effect sizes and their variances is straightforward using standard formulas (Radua et al., 2012). Meta-analysts can easily convert t-, z- and p-values from one to another with SDM “imgcalc”, or other similar tools.

The scenario is different when SDM imputes the effect-sizes images from the reported peak coordinates and t-values. The raw information is then scarce and SDM still has to recreate the 3D study images. To this end, AES-SDM first converts the reported t-values of the peaks into effect sizes using the formulas above. Starting from these “safe” points, it then imputes the effect sizes of the voxels surrounding the peaks as only slightly lower effect sizes than the effect sizes of the peaks. Afterwards, it imputes the effect sizes of the voxels surrounding these small blobs of voxels again as only slightly lower effect sizes than the voxels surrounding the peaks. And so on, until it reaches voxels too far from any

peak and imputes their effect size as null.

These imputations of AES-SDM are inexact, especially in the voxels further from peaks. To overcome this issue, SDM-PSI conducts multiple imputation following the PSI general conditions. Specifically, it uses AES-SDM kernels to estimate the lower and upper bounds of possible effect sizes for each study separately. Second, it uses MetaNSUE (Radua et al., 2015; Albajes-Eizagirre et al., 2018) (available at R CRAN and at <https://www.metansue.com/>) to estimate the most likely effect size and its standard error and create several imputations based on these estimations and the bounds. MetaNSUE is a method for univariate meta-analysis developed to include studies from which the meta-analytic researcher knows that the analysis was not statistically significant, but he/she cannot know the actual effect size (usually because authors of the study only wrote “n.s.”). Its empirical validation showed that this method is substantially less biased than assuming that the effect size is null (Radua et al., 2015), and the method has been recently improved to be robust in two scenarios frequent in CBMA, namely the scarcity of known data and the use of potential presence of very high t-values (e.g., $9.9 < z < 10$) (Albajes-Eizagirre et al., 2018). To adapt MetaNSUE for voxel-based neuroimaging, we had to ensure that the imputed images have a realistic local spatial structure (i.e., correlations between adjacent voxels are positive) but, importantly, the creation of this structure is not to the detriment of the accuracy of the imputations.

In any case, given the relevance of the peak coordinates and t-values in these recreations, the meta-analysts must extract them carefully from the paper, ensuring that the authors of the paper reported peaks from all the space of interest (e.g., the gray matter) and that they applied the same statistical threshold to all voxels (e.g., avoiding small volume corrections). Again, some studies may report z-values or p-values instead of t-values, but the meta-analysts may convert them in the <https://www.sdmproject.com/> website or using any other statistical converter.

3.2.1. Estimation of the lower and upper effect-size bounds

As a “pre-processing” step, SDM-PSI calculates an image of the lower bound of the possible effect sizes (i.e., the lowest potential effect size of each voxel) and an image of their upper bound (i.e., the largest potential effect size of each voxel) for each study.

The lower and upper effect-size bounds are obvious in a peak: both are the effect size of the peak. They are also relatively obvious in voxels

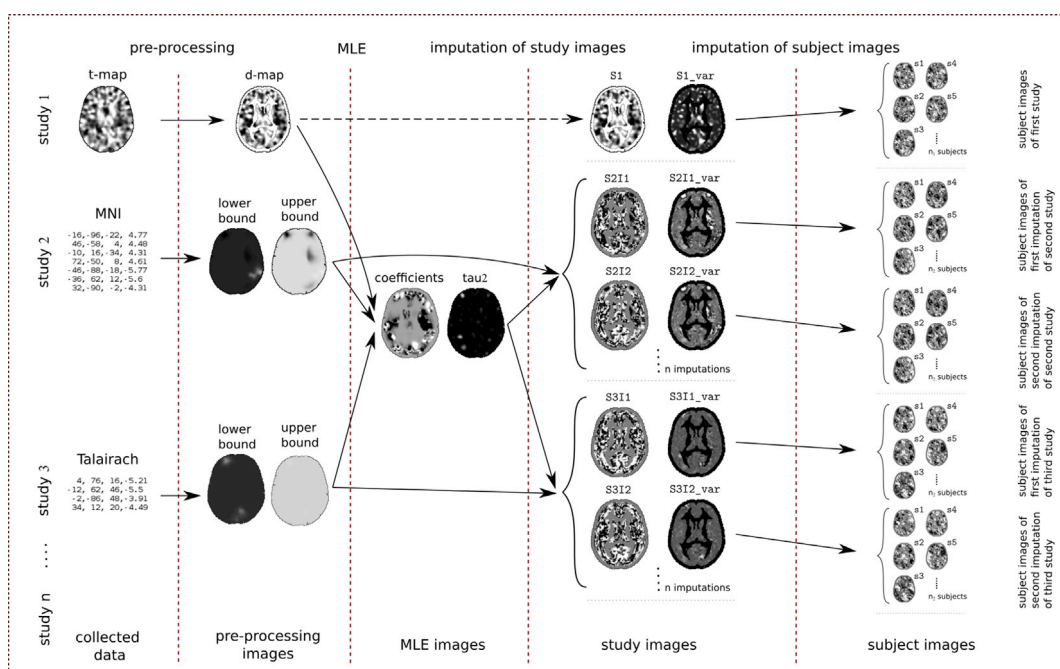


Fig. 3. PSI-SDM steps to impute subject images from collected data. *Footnote:* MLE: maximum likelihood estimation; MNI: Montreal Neurological Institute.

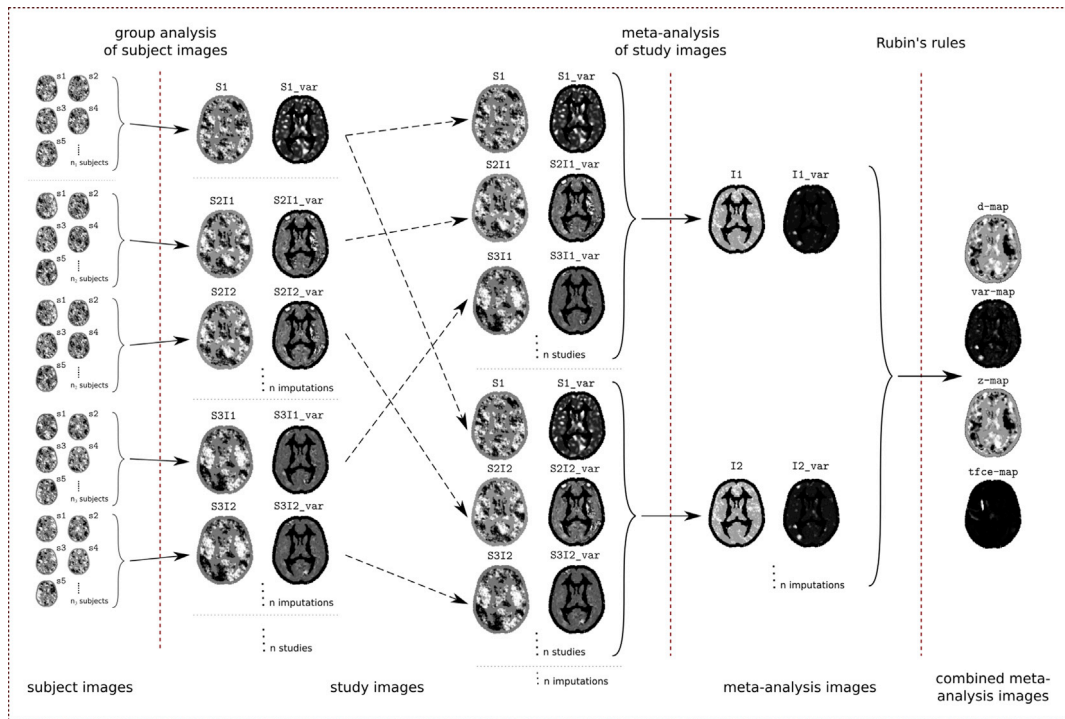


Fig. 4. PSI-SDM steps to combine subject images from the different imputations in a single combined meta-analysis image.

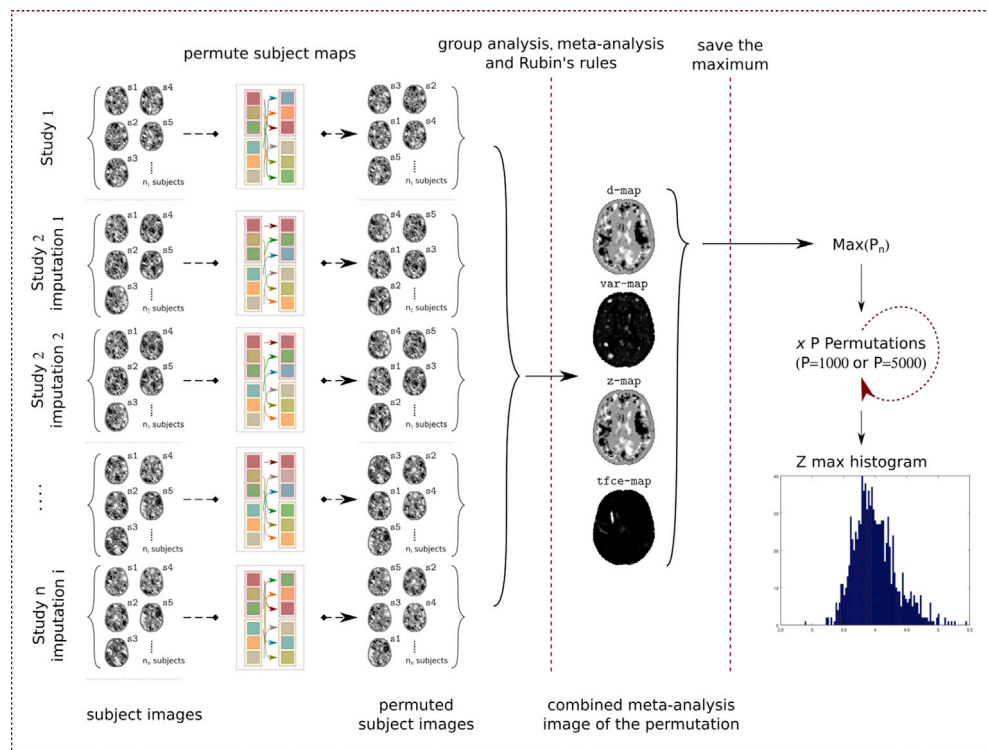


Fig. 5. PSI-SDM steps to conduct the subject-based permutation test.

far from any peak: they correspond to the positive and negative thresholds of statistical significance, because otherwise the studies would have found these voxels statistically significant.

Conversely, the procedure to establish effect-size bounds is more complex in voxels close to a peak, given that they should have effect-sizes similar to but lower than that of the peak, and some of them could be beyond the thresholds of statistical significance (i.e., they could have

been part of the cluster). SDM-PSI draws the upper effect-size bound as a descending smooth line from the effect size of the peak to the effect size of the positive threshold of statistical significance. Similarly, it draws the lower effect-size bound as a descending smooth line from the effect size of the peak to the effect size of the negative threshold of statistical significance. See a simplified version of these curves in Fig. 6.

More specifically, SDM-PSI uses the AES-SDM anisotropic Gaussian

kernels, which adapt the descending lines to the irregular spatial irregularities of the brain. We based this adaptation on the spatial covariance between each pair of adjacent voxels, which should capture spatial irregularities such as the boundaries between regions and tissues (e.g., the correlation between adjacent voxels is strong in the middle of a region and weak in a tissue boundary) (Radua et al., 2014):

$$y_{lower} = y_{\alpha/2} + \exp\left(\frac{-D^2}{2 \cdot \sigma_{kernel}^2}\right) \cdot (y_{peak} - y_{\alpha/2})$$

$$y_{upper} = y_{1-\alpha/2} + \exp\left(\frac{-D^2}{2 \cdot \sigma_{kernel}^2}\right) \cdot (y_{peak} - y_{1-\alpha/2})$$

where y_{lower} and y_{upper} are the effect-size bounds of the voxel of interest, y_{peak} is the effect size of the close peak, $y_{\alpha/2}$ and $y_{1-\alpha/2}$ are effect sizes of the thresholds of statistical significance, σ_{kernel} is the user-selected sigma of the kernel, and D is a virtual distance which depends on the real distance between the voxel and the peak (D_{real}), the correlation between the voxel and the peak (ρ) and the user-selected degree of anisotropy (α) (please see recommendations on these parameters in the Discussion):

$$D = \sqrt{(1 - \alpha) \cdot D_{real}^2 + \alpha \cdot 2\sigma_{kernel}^2 \cdot \log(\rho^{-1})}$$

SDM reads a theoretical correlation between each pair of adjacent voxels in a template, and it estimates the correlation between two non-adjacent voxels using a Dijkstra's algorithm (Radua et al., 2014). When a voxel is close to more than one peak, it conducts a weighted average of the effect sizes estimated from being close to each peak (Radua et al., 2014).

Of course, this step is unnecessary for studies from which the raw study image is available.

3.2.2. MLE of the effect size and its standard error

The first step after the pre-processing is the estimation of the most likely effect size and its standard error. In the simplest case of meta-analysis, this effect size is the same for all studies, whereas in a meta-regression and other analyses using linear models, the effect size of each study may depend on one or more covariates.

As in the work of Costafreda (Costafreda, 2012; Tobin, 1958; Schnedler, 2005), SDM-PSI estimates the parameters using maximum likelihood techniques. However, SDM-PSI adds several adjustments to prevent that a single or few studies drive the meta-analysis. These adjustments are required for a correct control of the FWER, and they are

already part of MetaNSUE (Albajes-Eizagirre et al., 2018). In any case, we must highlight that SDM-PSI only uses MLE as a starting point for the subsequent multiple imputation, avoiding the biases associated with single imputation (Rubin, 1987) and capturing the “uncertainty” of the unknown effect sizes as variance between imputations.

Relevantly, the likelihood to maximize for each study is not the likelihood of a specific effect size but the likelihood that the unreported effect size lays within the two effect size bounds. As detailed in (Radua et al., 2015) and (Albajes-Eizagirre et al., 2018), this likelihood is simply the difference of the cumulative normal distribution function evaluated at the upper and lower effect-size bounds:

$$L = \prod_{i=1}^N \left(\Phi\left(\frac{y_{upper,i} - X_i \cdot \beta}{\sqrt{v_{upper,i} + \tau^2}}\right) - \Phi\left(\frac{y_{lower,i} - X_i \cdot \beta}{\sqrt{v_{lower,i} + \tau^2}}\right) \right)$$

Note that when SDM-PSI knows the effect size of a study (e.g., because there is a peak in that voxel, or because the raw study image is available), the likelihood for this study is simply the probability function of the normal distribution.

The adjustments to prevent that a single or few studies drive the meta-analysis are similar to a trimmed mean: SDM-PSI conducts several MLE iterations that progressively discard the studies that increase the most the absolute MLE. These adjustments have little effects in voxels where the effect sizes are mostly known, whilst they prevent that a single or few studies drive the meta-analysis in voxels where the effect sizes are mostly unknown (Albajes-Eizagirre et al., 2018). Specifically, SDM-PSI conducts the estimation with all studies but the first, then with all studies but the second, then with all studies but the third, and so on, and only uses the combination returning the lowest absolute MLE. If the number of studies is large, this iteration is repeated to exclude a second study, and repeated again to exclude a third study, until the probability of a false positive meta-analytic effect size is not higher than 0.05, even in the worst-case scenario (Albajes-Eizagirre et al., 2018).

3.2.3. Multiple imputation

This step, separately conducted for each study, consists in imputing many times study images that meet the general PSI conditions adapted to SDM: a) the effect sizes imputed for a voxel must follow a truncated normal distribution with the MLE estimates and the effect-size bounds as parameters; and b) the effect sizes of adjacent voxels must show positive correlations.

An elegant solution to meet both conditions is unfortunately not straightforward, but SDM-PSI takes a pragmatic approach that, at the end of the day, yields imputed images that meet the two conditions. First, it assigns each voxel a uniformly distributed value between zero and one. Second, and separately for each voxel, it applies a threshold, spatial smoothing and scaling that ensures that the voxel has the expected value and variance of the truncated normal distribution and, simultaneously, has strong correlations with the neighboring voxels.

To ensure that the voxel has the expected value of the truncated normal distribution, the threshold applied to the voxels laying within the smoothing kernel is the expected value of the truncated normal distribution scaled to 0–1, and the number (between 0 and 1) resulting from the smoothing is rescaled to the bounds of the truncated normal distribution. To ensure that the voxel has the expected variance of the truncated normal distribution, SDM-PSI selects an anisotropic smoothing kernel that follows the spatial covariance of the voxel and makes the variance of the resulting value in the voxel coincide with that variance of the truncated normal distribution. Please note that each voxel must follow a different truncated normal distribution, and thus this thresholding/smoothing/rescaling process is different for each voxel.

Of course, this step is again unnecessary for studies from which the raw study image is available. It is neither conducted in peaks and in those voxels where the lower and upper effect-size bound are very close (e.g., difference < 0.02), because the simple mean of the effect-size bounds is already accurate.

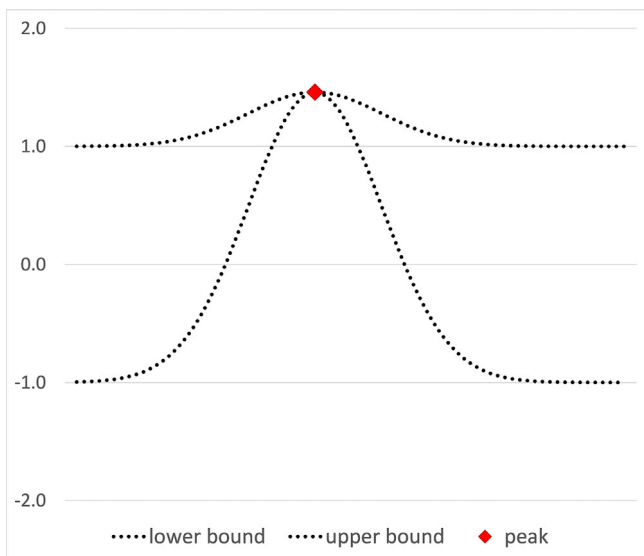


Fig. 6. PSI-SDM estimation of the lower and upper effect-size bounds for studies without raw study image available.

3.3. Imputation of subject images

For simplicity, the imputation function in SDM-PSI is a generation of random normal numbers with their mean equal to the sample effect size of the voxel in the (unpermuted) study image and unit variance. Note that the sample effect size is the effect size of the study after removing (i.e., dividing by) the J Hedge correction factor (Hedges and Olkin, 1985).

However, as explained earlier, the values of a voxel must show a Pearson correlation of one between any two imputed datasets, and the values of adjacent voxels must show realistic correlations observed in real humans. To meet these conditions, SDM-PSI only imputes a single, common preliminary dataset of subject images for all imputed datasets, and afterwards it scales it to the study image of each imputed dataset. In the common preliminary set, the values of any voxel have null mean, and adjacent voxels show the expected correlations. For instance, if the correlation observed in humans between voxels A and B is 0.67, the correlation between these two voxels in the imputed subject images must be 0.67. SDM-PSI uses the correlation templates created for AES-SDM (Radua et al., 2014) to know the correlation between every pair of two voxels (i.e., to take the irregular spatial covariance of the brain into account), but other approaches are possible. Afterwards, and separately for each imputed dataset, SDM-PSI simply adds the sample effect size of each voxel of the study image to all subject images. The complex part is thus

$$\begin{aligned}
 w_A &= r_{AY} - w_B r_{AB} - w_C r_{AC} - w_R r_{AR} \\
 w_B &= \frac{(r_{BY} - r_{AB} r_{AY}) - w_C (r_{BC} - r_{AB} r_{AC}) - w_R (r_{BR} - r_{AB} r_{AR})}{1 - r_{AB}^2} \\
 w_C &= \frac{\left([(1 - r_{AB}^2) r_{CY} - r_{AC} (r_{AY} - r_{AB} r_{BY}) - r_{BC} (r_{BY} - r_{AB} r_{AY})] \right. \\
 &\quad \left. - w_R [(1 - r_{AB}^2) r_{CR} - r_{AC} (r_{AR} - r_{AB} r_{BR}) - r_{BC} (r_{BR} - r_{AB} r_{AR})] \right)}{1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC}} \\
 w_R &= \sqrt{\frac{\left(\begin{aligned} &1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC} - (1 - r_{BC}^2) r_{AY}^2 - (1 - r_{AC}^2) r_{BY}^2 - (1 - r_{AB}^2) r_{CY}^2 \\ &+ 2(r_{AB} - r_{AC} r_{BC}) r_{AY} r_{BY} + 2(r_{AC} - r_{AB} r_{BC}) r_{AY} r_{CY} + 2(r_{BC} - r_{AB} r_{AC}) r_{BY} r_{CY} \end{aligned} \right)}{\left(\begin{aligned} &1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC} - (1 - r_{BC}^2) r_{AR}^2 - (1 - r_{AC}^2) r_{BR}^2 - (1 - r_{AB}^2) r_{CR}^2 \\ &+ 2(r_{AB} - r_{AC} r_{BC}) r_{AR} r_{BR} + 2(r_{AC} - r_{AB} r_{BC}) r_{AR} r_{CR} + 2(r_{BC} - r_{AB} r_{AC}) r_{BR} r_{CR} \end{aligned} \right)}}
 \end{aligned}$$

the creation of a common preliminary dataset of subject images that shows the expected correlation. We explain how SDM-PSI conducts it for one-sample studies in the following:

For imputing subject values (Y) in a voxel that has no neighboring voxels imputed yet, SDM-PSI simply creates random normal values and standardizes them to have null mean and unit variance (R):

$$Y = R$$

For imputing subject values in a voxel that has one neighboring voxel already imputed, SDM-PSI conducts a weighted average of the subject values of the neighboring voxel (A) and new standardized random normal values:

$$Y = w_A A + w_R R$$

where w are the weights that ensure that the resulting subject values have unit variance and the desired correlation (see mathematical derivation in the Supplement):

$$w_A = r_{AY} - w_R r_{AR} w_R = \sqrt{\frac{1 - r_{AY}^2}{1 - r_{AR}^2}}$$

For imputing subject values in a voxel that has two neighboring voxels already imputed, SDM-PSI conducts again a weighted average of

the subject values of the neighboring voxels (A and B) and new standardized random normal values:

$$Y = w_A A + w_B B + w_R R$$

where again w are the weights that ensure that the resulting subject values have unit variance and the desired correlations:

$$\begin{aligned}
 w_A &= r_{AY} - w_B r_{AB} - w_R r_{AR} \\
 w_B &= \frac{(r_{BY} - r_{AB} r_{AY}) - w_R (r_{BR} - r_{AB} r_{AR})}{1 - r_{AB}^2} \\
 w_R &= \sqrt{\frac{1 - r_{AB}^2 - r_{AR}^2 - r_{BR}^2 + 2r_{AB} r_{AR} r_{BR}}{1 - r_{AB}^2 - r_{AR}^2 - r_{BR}^2 + 2r_{AB} r_{AR} r_{BR}}}
 \end{aligned}$$

Finally, for imputing subject values in a voxel that has three neighboring voxels already imputed, SDM-PSI conducts once more a weighted average of the subject values of the neighboring voxels (A , B and C) and new standardized random normal values:

$$Y = w_A A + w_B B + w_C C + w_R R$$

where w are once more the weights that ensure that the resulting subject values have unit variance and the desired correlations:

Note that as far as the imputation of the voxels follows a simple order and the software only accounts for correlations between voxels sharing a face, a voxel cannot have more than three neighbor voxels already imputed. For example, imagine that the imputation follows a left/posterior/inferior to right/anterior/superior direction. When the software imputes a given voxel, it will have already imputed the three neighbors in the left, behind and below, while it will impute later the three neighbors in the right, in front and above. The number of neighbor voxels imputed or to impute will be lower if some of them are outside the mask.

For two-sample studies, SDM-PSI imputes subject values separately for each sample, and it only adds the effect size to the patient (or non-control) subject images.

3.4. Permutations

The permutation algorithms are general.

3.5. Group analysis, meta-analysis and Rubin's rules

In SDM-PSI, the group analysis is the estimation of Hedge-corrected effect sizes. In practice, this estimation simply consists of calculating the mean (or the difference of means in two-sample studies) and multiplying by J , given that imputed subject values have unit variance.

The meta-analysis consists of the fitting of a standard random-effects

model. The design matrix includes any covariate used in the MLE step, and the weight of a study is the inverse of the sum of its variance and the between-study heterogeneity τ^2 , which in SDM-PSI may be estimated using either the DerSimonian-Laird or the slightly more accurate restricted-maximum likelihood (REML) method (Viechtbauer, 2005; Thorlund et al., 2011). After fitting the model, SDM conducts a standard linear hypothesis contrast and derives standard heterogeneity statistics H^2 , I^2 and Q .

Finally, SDM-PSI uses Rubin's rules to combine the coefficients of the model, their covariance and the heterogeneity statistics I and Q of the different imputed datasets (Li et al., 1991; Radua et al., 2015; Albajes-Eizaguirre et al., 2018). Note that Q follows a χ^2 distribution, but its combined statistic follows an F distribution. For convenience, SDM-PSI converts F_Q back into a Q (i.e. converts an F statistic to a χ^2 statistic with the same p-value). It also derives $H_{combined}$ from $I_{combined}$.

3.6. Maximum statistic test

SDM-PSI can currently save four different maximum statistics from the image of z-values: the largest z-value (i.e., voxel-based statistics), the maximum cluster size or mass after thresholding with a user-defined z-value (i.e., cluster-size or mass statistics) (Bullmore et al., 1999), and the maximum TFCE (Smith and Nichols, 2009). We have implemented TFCE in our software so that users who do not have FSL installed will still be able to use TFCE in SDM-PSI.

This procedure theoretically only tests hypothesis in one direction (e.g., patients > controls), but not the hypothesis in the other direction (e.g., patients < controls), and thus, the procedure should be conducted twice, one for each direction. However, given that the hypotheses are complementary and a permutation test is computationally consuming, SDM-PSI saves two numbers from each permutation: one for the positive hypothesis (e.g., the highest z-value) and one for the negative hypothesis (e.g., the lowest z-value), in two separate null distributions.

3.7. Meta-regression and other linear models

As stated earlier, permutations for a meta-regression and other linear models can only be at the study-level, because we only know the value of the moderators at this level. Several permutation approaches are possible, but SDM-PSI uses the Freedman-Lane procedure for its optimal statistical properties (Winkler et al., 2014).

4. Validation of SDM-PSI

4.1. Control of the FWER

We checked empirically whether the SDM-PSI controls the FWER at the desired level. Specifically, we conducted hundreds of meta-analyses of (simulated) studies comparing the gray matter volume of random groups of subjects, and thresholded them to control the FWER at 5%. We expected that only 5% of these meta-analyses would return one or more (false positive) findings.

We used 1158 real brain structural MR images to simulate the studies. We had already acquired them for previous studies of the unit, and refer the reader to the corresponding manuscripts for details of the acquisition and pre-processing steps (Amann et al., 2016; Moreno-Alcazar et al., 2016; Vicens et al., 2016; Landin-Romero et al., 2016). Independently for each simulated meta-analysis, we randomly divided the 1158 sound MNI-registered images into several (simulated) studies with varying sample sizes. Small studies included 22, 26, 32, 40 or 48 subjects, and large studies included 60, 72, 88, 108 or 134 subjects. We chose these sample sizes because they follow a plausible exponential distribution (i.e., a meta-analysis commonly includes more small studies than large studies) and are common in neuroimaging studies (total participants = 22–48 for small studies, 60–134 for large studies). Small meta-analyses included 10 studies and large meta-analyses included 20

studies. We conducted 400 small meta-analyses of small studies, 400 small meta-analyses of both small and large studies, 400 large meta-analyses of small studies, and 400 large meta-analyses of both small and large studies.

For each simulated study, we first conducted a t -test to detect gray matter differences between the simulated patients and controls, we then thresholded the resulting t -value image with $p < 0.001$ uncorrected (for both patients > controls and patients < controls), and finally saved the peaks of the clusters of statistical significance, miming how they are commonly reported in published neuroimaging studies. Afterwards, we conducted a simulated meta-analysis with SDM-PSI exclusively using the coordinates and t -values of the simulated studies. Here we did not use any raw study images, which would substantially increase the accuracy of the meta-analysis, because we wanted to test the performance of SDM-PSI under the more challenging, only-peaks scenario. Please see the next part of the validation for simulations including raw study images.

We thresholded each simulated SDM-PSI meta-analysis using voxel, cluster-size, cluster-mass and TFCE statistics. For cluster-size and mass, we used the z -thresholds 2.33 and 3.09, corresponding to uncorrected $p = 0.01$ and 0.001 . For TFCE, we used FSL default parameters: extension power $E = 2$ and height power $H = 0.5$ (Smith and Nichols, 2009). Given that the simulated studies compared random groups of subjects, we calculated the empirical FWER as the percentage of simulated meta-analyses with one or more findings, and estimated its 95% confidence interval using the Clopper and Pearson exact method (Clopper and Pearson, 1934).

As we show in Table 1, SDM-PSI globally controlled the FWER below 5% for in all scenarios except for one: when we applied cluster-based statistics with high z -thresholds in small meta-analyses of small studies (FWER increased to 15–23%). However, the control was too conservative in most scenarios, namely those involving the use of voxel-based statistics or the inclusion of many and/or large studies (FWER decreased to 0–4%).

4.2. Comparison with a pooled analysis of all subject data

As a proof of concept, we also compared the results of a SDM-PSI meta-analysis with the results of a standard SPM analysis of all the raw subject images of the studies included in a meta-analysis. The reader can find details of these functional magnetic resonance imaging (fMRI) data in (Radua et al., 2012). Briefly, the meta-analysis included 10 studies of the brain response to the presentation of fearful faces. We conducted 11 meta-analyses: one with only peak information, one with 1 raw study image, one with 2 raw study images, and so on until we conducted one with only raw study images. For comparison purposes, we also conducted the 11 meta-analyses with AES-SDM. We used default thresholds (SPM: voxel-based FWER < 0.05; SDM-PSI: TFCE-based FWER < 0.05; AES-SDM: voxel-based uncorrected $p < 0.005$ with peak SDM- $Z > 1$). We discarded clusters smaller than 10 voxels in all cases. The statistics of interest for each meta-analysis were the cluster-based sensitivity (i.e. the proportion of SPM activation clusters detected by the meta-analysis), the voxel-based sensitivity (i.e., the proportion of SPM activated voxels also labeled as activated by the meta-analysis), the voxel-based specificity (i.e., the proportion of SPM non-activated voxels also labeled as non-activated by the meta-analysis), the voxel-based accuracy (i.e., the proportion of SPM voxels correctly labeled as activated or non-activated by the meta-analysis), and computation time with a machine equipped with an Intel Xeon E5-2680@2.40Ghz processor and 128GB of RAM.

Cluster-based sensitivity was 100% for both SDM-PSI and AES-SDM in all cases. Voxel-based sensitivity was 38% for SDM-PSI and 65% for AES-SDM in the meta-analysis conducted with only peak information. For SDM-PSI, it increased to 98% with 1–2 raw study images, and to 100% with ≥ 3 raw study images. For AES-SDM, it increased to 84% with 1 study image, 93% with 2, 98% with 3, and 100% with ≥ 4 raw study images. Voxel-based specificity and accuracy were >92% in all cases. Computation time for SDM-PSI employing 50 threads was 56 min for the meta-analysis conducted with only peak information and 37 min for the

meta-analysis conducted with only raw study images. Computation time for AES-SDM was 4 min in both cases.

5. Discussion

This paper reports a novel algorithm for CBMA that, as opposed to current CBMA methods, conducts a standard subject-based permutation test to control the FWER. We have implemented and validated the method for SDM, but other developers might implement it for other CBMA methods. The software is freely available at <https://www.sdmproject.com/>. The clear strength of the new algorithm, to which we refer as PSI, is the use of standard statistical procedures, which avoid the drawbacks of the alternative procedures used in current CBMA methods (Albajes-Eizagirre and Radua, 2018).

Regarding the selection of parameters, we generally recommend the use of full anisotropy during the imputation of study images, and the use of TFCE in the statistical thresholding. On the one hand, in the validation of AES-SDM, we found that full anisotropy yielded relatively accurate estimations, and that this accuracy does not depend on the size of the kernel used (Radua et al., 2014). Alternatively, the meta-analyst may estimate the optimal imputation parameters for each meta-analysis. For example, in a previous meta-analysis in which we had many raw study images, we recreated these images using the peak information reported in the respective papers, using many combinations of parameters, and selected the combination of parameters that best recreated the images. We then used this combination for recreating the images of the studies from which we only had peak information (Fullana et al., 2016). On the other hand, the validation of SDM-PSI showed that voxel-based statistics might be too strict, cluster-based statistics may be too liberal in some scenarios, and TFCE statistics were neither too conservative nor too liberal.

Even if an aim of SDM-PSI is to impute the non-reported effect sizes with as much accuracy as possible, we suggest that researchers understand the imputed study images as a low-quality version of the raw study images. Indeed, in the second part of the validation we found that sensitivity increased from 38% to 98% with the inclusion of a single raw study image. Even if we suggest that the readers take this part of the validation with some caution because it is only a proof of concept example, we can safely say that a meta-analysis should improve if the meta-analysts are able to include raw study images instead of peak coordinates for some (or all) studies. At this regard, we have to mention that there exist several great data sharing initiatives, such as NeuroVault (Gorgolewski et al., 2015), that allow an easy storage and sharing of raw study images.

As we noted in the Introduction, a sign-flipping study-based permutation would be quicker and would similarly test whether effects are not null. This approach would have steps similar to PSI but it would directly permute study images and thus would not need to impute subject images, permute them, and conduct the group analysis of the permuted subject images. However, the software imputes the subject images only once, and their permutation and group analysis are proportionally very quick, for what the decrease in computation time would be small. The computationally demanding steps are others, such as the meta-analysis. Moreover, for meta-analyses with a small number of studies, the estimation of the p-value would be poorer. We provide in the Supplement a script in R-language to compare execution time and accuracy of study- and subject-based permutation of a single variable. With 10 studies, the study-based permutation is 23% quicker than the subject-based permutation, but the mean squared error in the estimation of the p-values is 107% larger. The gain in computation time would be proportionally smaller in SDM-PSI, because there are other steps that the software would still need to do, such as the spatial statistics.

With few exceptions (see below), SDM-PSI controlled the FWER below 5%, but it was too conservative. This means that in the absence of true effects, an SDM-PSI meta-analysis should rarely detect false effects, but in the presence of weak but true effects, an SDM-PSI meta-analysis

may fail to detect them. Conversely, the control failed (it was too liberal) when we applied cluster-based statistics with high z-thresholds in small meta-analysis of small studies. This liberal behavior is intriguing. On the one hand, an influential previous study reported increased positive rates with the use of cluster-based statistics in fMRI (Eklund et al., 2016). While the authors mostly found these increases in parametric statistical tests, the permutation tests were not entirely free from problems. On the other hand, we also suspect that a sum of slight inaccuracies may increase the FWER in cluster-based statistics. First, we have already noted that despite our efforts, the imputed images are not perfect recreations of the raw images. Thus, we expect some error in the structures of the spatial covariance of the imputed images, and this error may distort, to a small extent, the clustering of statistically significant voxels. Second, it is known that the estimation of between study heterogeneity may be less accurate in meta-analyses with few studies. An underestimation could decrease the variability between imputed images of the same study, increasing the false positive rate.

The validation showed that AES-SDM might be more voxel-based sensitive than SDM-PSI when only peak information is available. We already suggested that the readers take this part of the validation with some caution because it is only a proof of concept example, but it is plausible that AES-SDM is indeed more sensitive than SDM-PSI when only peak information is available. On the one hand, the former conducts a test for convergence, which may be low-powered when there are multiple true effects, but may be high-powered when there is only one or two (Albajes-Eizagirre and Radua, 2018). On the other hand, we developed the latter with a deep focus on the control of the FWER, including the leave-one-out protection in the MLE step, and this may be associated with conservative p-values. In any case, we would like to highlight that both SDM-PSI and AES-SMD showed a remarkable voxel-based specificity and accuracy even in the absence of raw study images.

SDM-PSI has some limitations. First, studies correcting for multiple comparisons may not report the t-threshold, or they may do not even have used one (e.g., if they used a TFCE threshold). For these cases, we recommend using the t-threshold equivalent to $p = 0.001$ uncorrected, which is a conservative choice because if the study had used this threshold it would very likely have found a large number of statistically significant voxels that will be erroneously considered non-statistically significant by SDM-PSI. Second, SDM-PSI assumes that all voxels far from any peak were non-statistically significant. However, there is the possibility that some isolated voxels of the raw study image may have reached statistical significance but were unreported due to small cluster extent. However, this possibility should have a limited conservative impact on the meta-analysis. Third, the new method downwards biases the effect sizes and z-values due to the adjustments to prevent that a single or few studies drive the meta-analysis in the MLE step. This means that uncorrected p-values directly derived from the z-values may be slightly conservative. However, we believe that this prevention is more important than the resulting conservative bias because it prevents that findings from a single or few studies have an erroneously large influence on the meta-analysis. Fourth, the spatial structure of the imputed images is realistic and based on anisotropic correlation templates, but it may still be different from that of the raw studies. We expect, though, that the differences should be substantially milder than when assuming an isotropic brain, and the validation showed a good control of the FWER with few exceptions. Fifth, the estimation of between-study heterogeneity may be less accurate in meta-analyses with few studies. In SDM-PSI, this effect could affect the meta-analysis (i.e., as in any other meta-analytic method), but it could also affect the multiple imputation of study images (see above). Sixth, during the imputation of subject images, the new software only accounts for correlations between voxels sharing a face. Thus, during this step it considers that a voxel has only six neighbors (one in the left, one in the right, one behind, one in front, one below, and one above). The software may be rewritten to also account for correlations between voxels sharing only an edge or only a vertex, but we believe that these calculations may take an important computational

effort while provide little increase in accuracy. Finally, SDM-PSI is substantially more computationally demanding than AES-SDM.

Conflicts of interest

The author reports no conflicts of interests related to this manuscript.

Acknowledgements

This work was supported by PFIS Predoctoral Contract FI16/00311, Miguel Servet Research Contract MS14/00041 and Research Project PI14/00292 from the Plan Nacional de I + D + i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund (FEDER). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neuroimage.2018.10.077>.

References

- Albajes-Eizaguirre, A., Radua, J., 2018. What do results from coordinate-based meta-analyses tell us? *Neuroimage*.
- Albajes-Eizaguirre, A., Solanes, A., Radua, J., 2018. Meta-analysis of non-statistically significant unreported effects (MetaNSUE). *Stat. Methods Med. Res.* (in press).
- Amann, B.L., et al., 2016. Brain structural changes in schizoaffective disorder compared to schizophrenia and bipolar disorder. *Acta Psychiatr. Scand.* 133 (1), 23–33.
- Bossier, H., et al., 2017. The influence of study-level inference models and study set size on coordinate-based fMRI meta-analyses. *Front. Neurosci.* 11, 745.
- Bullmore, E.T., et al., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.* 18 (1), 32–42.
- Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413.
- Costafreda, S.G., 2012. Parametric coordinate-based meta-analysis: valid effect size meta-analysis of studies with differing statistical thresholds. *J. Neurosci. Methods* 210 (2), 291–300.
- Costafreda, S.G., David, A.S., Brammer, M.J., 2009. A parametric approach to voxel-based meta-analysis. *Neuroimage* 46 (1), 115–122.
- Eickhoff, S.B., et al., 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30 (9), 2907–2926.
- Eickhoff, S.B., et al., 2012. Activation likelihood estimation meta-analysis revisited. *Neuroimage* 59 (3), 2349–2361.
- Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.* 113 (28), 7900–7905.
- Fullana, M.A., et al., 2016. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Mol. Psychiatr.* 21 (4), 500–508.
- Gorgolewski, K.J., et al., 2015. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinf.* 9, 8.
- Hedges, L.V., Olkin, I., 1985. *Statistical Methods for Meta-analysis*. Academic Press, Orlando.
- Holmes, A.P., et al., 1996. Nonparametric analysis of statistic images from functional mapping experiments. *J. Cerebr. Blood Flow Metabol.* 16 (1), 7–22.
- Kang, J., et al., 2011. Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *J. Am. Stat. Assoc.* 106 (493), 124–134.
- Kang, J., et al., 2014. A Bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *Ann. Appl. Stat.* 8 (3), 1800–1824.
- Laird, A.R., et al., 2005. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25 (1), 155–164.
- Landin-Romero, R., et al., 2016. Midline brain abnormalities across psychotic and mood disorders. *Schizophr. Bull.* 42 (1), 229–238.
- Li, K.H., et al., 1991. Significance levels from repeated p-values with multiply-imputed data. *Stat. Sin.* 1 (1), 65–92.
- Montagna, S., et al., 2017. Spatial Bayesian Latent Factor Regression Modeling of Coordinate-based Meta-analysis Data. *Biometrics*.
- Moreno-Alcazar, A., et al., 2016. Brain abnormalities in adults with Attention Deficit Hyperactivity Disorder revealed by voxel-based morphometry. *Psychiatr. Res.* 254, 41–47.
- Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15 (1), 1–25.
- Phipson, B., Smyth, G.K., 2010. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* 9, Article 39.
- Radua, J., Mataix-Cols, D., 2009. Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *Br. J. Psychiatry* 195 (5), 393–402.
- Radua, J., Mataix-Cols, D., 2012. Meta-analytic methods for neuroimaging data explained. *Biol. Mood Anxiety Disord.* 2, 6.
- Radua, J., et al., 2012. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *Eur. Psychiatr.* 27 (8), 605–611.
- Radua, J., et al., 2014. Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Front. Psychiatr.* 5, 13.
- Radua, J., et al., 2015. Ventral striatal activation during reward processing in psychosis: a neurofunctional meta-analysis. *JAMA Psychiatry* 72 (12), 1243–1251.
- Rubin, D.B., 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.
- Schneider, W., 2005. Likelihood estimation for censored random vectors. *Econom. Rev.* 24 (2), 195–217.
- Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44 (1), 83–98.
- Thorlund, K., et al., 2011. Comparison of statistical inferences from the DerSimonian-Laird and alternative random-effects model meta-analyses - an empirical assessment of 920 Cochrane primary outcome meta-analyses. *Res. Synth. Methods* 2 (4), 238–253.
- Tobin, J., 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26 (1), 24–36.
- Turkeltaub, P.E., et al., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16 (3 Pt 1), 765–780.
- Turkeltaub, P.E., et al., 2012. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum. Brain Mapp.* 33 (1), 1–13.
- Vicens, V., et al., 2016. Structural and functional brain changes in delusional disorder. *Br. J. Psychiatry* 208 (2), 153–159.
- Viechtbauer, W., 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 30 (3), 261–293.
- Wager, T.D., Lindquist, M., Kaplan, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cognit. Affect Neurosci.* 2 (2), 150–158.
- Winkler, A.M., et al., 2014. Permutation inference for the general linear model. *Neuroimage* 92, 381–397.
- Wise, T., et al., 2016. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. *Mol. Psychiatr.*
- Yue, Y.R., Lindquist, M.A., Loh, J.M., 2012. Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *Ann. Appl. Stat.* 6 (2), 697–718.

Video Article

Meta-analysis of Voxel-Based Neuroimaging Studies using Seed-based d Mapping with Permutation of Subject Images (SDM-PSI)

Anton Albajes-Eizagirre^{1,2}, Aleix Solanes^{1,2}, Miquel Angel Fullana^{2,3}, John P. A. Ioannidis⁴, Paolo Fusar-Poli^{5,6,7}, Carla Torrent^{1,2,3,8}, Brisa Solé^{1,2,3,8}, Caterina Mar Bonnín^{1,2,3,8}, Eduard Vieta^{1,2,3,8}, David Mataix-Cols⁹, Joaquim Radua^{1,2,5,9}

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

²Mental Health Research Networking Center (CIBERSAM)

³Institute of Neurosciences, Hospital Clinic de Barcelona

⁴Departments of Medicine, of Health Research and Policy, and of Biomedical Data Science, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences

⁵Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London

⁶OASIS Service, South London and Maudsley NHS Foundation Trust

⁷Department of Nervous System and Behavioral Sciences, University of Pavia

⁸University of Barcelona

⁹Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet

Correspondence to: Anton Albajes-Eizagirre at antonae@gmail.com, Joaquim Radua at radua@clinic.cat

URL: <https://www.jove.com/video/59841>

DOI: [doi:10.3791/59841](https://doi.org/10.3791/59841)

Keywords: Neuroscience, Issue 153, familywise error rate, functional magnetic resonance imaging (fMRI), meta-analysis, neuroimaging, permutation of subject images (PSI), seed-based d mapping (SDM), threshold-free cluster enhancement (TFCE), voxel-based morphometry (VBM)

Date Published: 11/27/2019

Citation: Albajes-Eizagirre, A., Solanes, A., Fullana, M.A., Ioannidis, J.P., Fusar-Poli, P., Torrent, C., Solé, B., Bonnín, C.M., Vieta, E., Mataix-Cols, D., Radua, J. Meta-analysis of Voxel-Based Neuroimaging Studies using Seed-based d Mapping with Permutation of Subject Images (SDM-PSI). *J. Vis. Exp.* (153), e59841, doi:10.3791/59841 (2019).

Abstract

Most methods for conducting meta-analysis of voxel-based neuroimaging studies do not assess whether effects are not null, but whether there is a convergence of peaks of statistical significance, and reduce the assessment of the evidence to a binary classification exclusively based on p-values (i.e., voxels can only be "statistically significant" or "non-statistically significant"). Here, we detail how to conduct a meta-analysis using Seed-based d Mapping with Permutation of Subject Images (SDM-PSI), a novel method that uses a standard permutation test to assess whether effects are not null. We also show how to grade the strength of the evidence according to a set of criteria that considers a range of statistical significance levels (from more liberal to more conservative), the amount of data or the detection of potential biases (e.g., small-study effect and excess of significance). To exemplify the procedure, we detail the conduction of a meta-analysis of voxel-based morphometry studies in obsessive-compulsive disorder, and we provide all the data already extracted from the manuscripts to allow the reader to replicate the meta-analysis easily. SDM-PSI can also be used for meta-analyses of functional magnetic resonance imaging, diffusion tensor imaging, position emission tomography and surface-based morphometry studies.

Video Link

The video component of this article can be found at <https://www.jove.com/video/59841/>

Introduction

Since the introduction of magnetic resonance imaging, the neuroimaging community has published thousands of studies of the neural substrates of psychological functions and neuropsychiatric disorders. To summarize these findings, several methods have been developed^{1,2,3,4,5,6}. Original voxel-based neuroimaging studies report the coordinates of the peaks of statistical significance (e.g., in a comparison of gray matter volume between patients and controls), and meta-analytic methods commonly assess whether there is convergence of peaks in certain brain regions.

However, we have previously shown that these tests for convergence of peaks rely on delicate assumptions that might influence the patterns of meta-analysis results and their statistical significance⁷. Specifically, these tests assume that voxels are independent and that they have the same probability of a "false" peak, while in real gray matter, voxels correlate with their neighbors and the probability that a voxel has a "false" peak depends on its tissue composition. In addition, they also encompass paradoxes such as that the statistical power increases in the presence of few true effects, and decreases when there are multiple true effects.

To overcome these problems, we developed a method that imputes the brain maps of statistical effects for each study and then conducts a standard random-effects meta-analysis to formally test whether the effects are different from zero. This method is called "Seed-based d Mapping with Permutation of Subject Images" (SDM-PSI)⁸ and its main features include:

- Accounting for both increases and decreases of the outcome of interest (e.g., activation and deactivation) so that contradictory findings cancel each other⁴;
- Use of effect size estimates with random-effects modeling, which increases reliability and performance⁹;
- Potential simultaneous inclusion of available 3D statistical images (i.e., maps of t-test values)¹⁰;
- Subject-based permutation tests identical to those of FSL "randomize" tool¹¹;
- Use of threshold-free cluster enhancement (TFCE) statistics¹².

We have detailed and fully validated the SDM methods elsewhere^{4,8,10,13,14}.

In addition, we suggest not relying on a binary classification of the voxels based on the level of statistical significance (significant vs. not significant) but, conversely, assessing the strength of the evidence using a set of criteria²². The binary statistical significance reductionism leads to poor control of the false positive and false negative rates¹⁵, whereas the criteria use ranges of statistical significance levels and take into account the amount of data or potential biases. The SDM-PSI software returns the necessary elements to conduct such a classification⁹ and thus they can be employed to afford a more granular classification of the strength of the evidence.

Here we show how to conduct a meta-analysis of voxel-based neuroimaging studies using SDM-PSI. To exemplify the protocol, we use data from a published meta-analysis of voxel-based morphometry studies that investigated gray matter abnormalities in patients with obsessive-compulsive disorder (OCD)⁴. However, we will not use the methods employed in that early meta-analysis, but the aforementioned state-of-the-art procedures. The reader can download the software and these data from our website (<http://www.sdmproject.com/>) to replicate the analysis.

All researchers who aim to conduct a meta-analysis of voxel-based neuroimaging studies can follow this protocol. The method can be used with functional magnetic resonance imaging (fMRI, e.g., BOLD response to a stimulus)¹⁶, voxel-based morphometry (VBM, e.g., gray matter volume)¹⁷, diffusion tensor imaging (DTI, e.g., fractional anisotropy)¹⁸, position emission tomography (PET, e.g., receptor occupancy)¹⁹ and surface-based morphometry (SBM, e.g. cortical thickness) studies/datasets.

Protocol

1. Installation of SDM-PSI

1. Go to <https://www.sdmproject.com/software/> to download the version of SDM-PSI for the operating system of the computer as a ZIP file.
2. Decompress the ZIP file. To avoid problems, decompress it within a local folder without blank spaces in its path.
3. Click the file **SdmPsiGui** to execute the graphical interface of SDM-PSI, and close the **About** splash window that will automatically open.
4. If **SdmPsiGui** does not find all required paths, it will automatically offer to display the preferences window. Press **Yes**.
 1. If MRICron is not installed in the computer go to https://www.nitrc.org/frs/?group_id=152 to download the version for the operating system of the computer as a ZIP file, and decompress the ZIP file.
 2. In the **Brain viewer** tab of the preferences window, ensure that **Brain viewer** is set to **MRICron**, and click the folder icon next to **Brain viewer executable** to find the MRICron executable.
 3. Ensure that all remaining paths in the different tabs have blue marks, which indicate that the paths are correct.
5. In case that SdmPsiGui has not automatically displayed the preferences window, go to the **Tools** menu and click **Preferences**.
6. In the **Multithreading** tab, specify the number of concurrent threads to use in the calculations. Some SDM-PSI calculations take a very long time (from hours to days) and consume a large amount of RAM memory (from hundreds of megabytes to gigabytes). The use of multiple threads (parallel processing) substantially decreases the time but increases the memory used.
7. Close the preferences window and SdmPsiGui.

2. Meta-analysis plan

1. Specify a precise question. For instance, "do patients with OCD have gray matter volume regional abnormalities?"
2. Write clear inclusion criteria that allow a systematic inclusion of the studies. For instance, "all studies that performed whole-brain voxel-based comparison of gray matter volume between individuals with OCD and healthy controls".
3. Write clear exclusion criteria that allow a systematic exclusion of those studies that cannot or should not be included for specific reasons. For instance, "studies with less than 10 patients, duplicated datasets, or studies from which the required information cannot be retrieved".
4. Write down the data to extract from each study. The following list includes the recommended data (some of them are not strictly required, but their absence would impoverish the meta-analysis):
 - An identification of the study.
 - The sample sizes.
 - The statistical significance level, which is the t-value, z-value or p-value used in the study to determine which voxels were statistically significant.
 - The software and stereotactic space. See **Table 1** for the list of software packages and stereotactic spaces understood by SDM-PSI.
 - The coordinates and height of the peaks. The height of a peak is its t-value or z-value, but a p-value is also useful.
 - Variables that will be used to describe the samples or to conduct subgroup analyses or meta-regressions.
5. To increase the quality of the review, consider following the "Ten simple rules for neuroimaging meta-analysis"²⁰ and the PRISMA checklist²¹.
6. To increase the transparency of the review, consider registering the protocol beforehand on a publicly available database such as PROSPERO (<https://www.crd.york.ac.uk/PROSPERO/>).

3. Exhaustive search

1. Select a set of keywords that allow finding any study that might meet the inclusion criteria. For instance, the keywords might be "obsessive-compulsive disorder" plus "morphometry", "voxel-based" or "voxelwise".
2. Conduct the search on databases such as PubMed and Web of Science:
 1. Go to database website, e.g. <https://www.ncbi.nlm.nih.gov/pubmed/> for PubMed.
 2. Type the search query. In the example meta-analysis, the query might be "obsessive-compulsive disorder" AND ("morphometry" OR "voxel-based" OR "voxelwise"). In this query, the operator "AND" means that the studies must have all keywords, the operator "OR" means that the studies must have at least one of the keywords, and the parentheses indicate the order of these logical operations. Therefore, the retrieved studies must have the keyword "obsessive-compulsive disorder" and at least one of the keywords "morphometry", "voxel-based" or "voxelwise". Note that other strategies are possible.
3. Apply the inclusion/exclusion criteria. For instance, from the results provided by the databases, select only the articles that analyze differences in gray matter volume between patients with obsessive-compulsive disorder and controls performing whole-brain voxel-based morphometry studies, and discard studies including less than 10 patients and studies that re-analyzed previously published data.
4. To increase the exhaustiveness of the search, consider conducting a manual search over the referenced works in the selected studies.
5. To maximize the inclusion of studies and avoid uncertainty in the collection of data, consider contacting the corresponding authors to ask for any missing or unclear data.
6. Record the number of studies retrieved and the number of studies excluded for each reason. Consider creating a PRISMA flow diagram²¹ with these numbers.

4. Collection of the data

1. For each included study, read the manuscript to find the specific data to extract.
2. Save the data from the studies systematically, e.g., typing the data in pre-formatted spreadsheet files. To minimize typing errors, consider copying and pasting the numbers of interest and double-checking the saved data.
3. When the statistical significance level is unclear, consider following these recommendations:
 1. If the manuscript reports peaks obtained using two whole-brain statistical significance levels, e.g. p-value < 0.001 without correction for multiple comparisons (from now on, "uncorrected threshold") and familywise error rate (FWER) < 0.05 (from now on "corrected threshold"), select the uncorrected threshold, and include all peaks obtained using this uncorrected threshold. The reason to prefer the uncorrected threshold is that studies usually obtain more peaks applying an uncorrected threshold, and SDM estimates the maps more accurately if it has information from more peaks.
 2. If the manuscript reports peaks obtained using an uncorrected threshold for increases and a corrected threshold for decreases (or vice versa), select the uncorrected threshold but include only the peaks obtained using the corrected threshold. This is a conservative approximation because it might discard some peaks obtained using the uncorrected threshold. An example of this situation might be when a manuscript states something such as "we detected FWER-corrected larger gray matter volumes in some regions, while we did not detect smaller gray matter volume in any region even using uncorrected p-value < 0.001".
 3. If the authors applied cluster-based statistics, use the cluster-forming height threshold. This is a conservative approximation because some voxels might have had t-values higher than the threshold, but the authors discarded them because their clusters were not large enough.
 4. If the manuscript does not specify a threshold, use a slightly smaller value than the t-value of the smallest peak. The reason to use this t-value is that if the authors had applied this statistical significance threshold without requiring a minimum size for the clusters, they would have found the same peaks.
4. When recording the peak information do the following:
 1. Exclude peaks obtained using a statistical significance threshold that is more liberal than the threshold selected for the rest of the brain. An example of this situation is when the authors applied more liberal thresholds or small volume corrections to some a priori brain regions.
 2. Convert z-values and p-values into t-values. Click the button **Convert peaks** in the SDM-PSI software to convert them easily. Alternatively, convert them in the same spreadsheet file (e.g., "=T.INV(1-0.001,34)" for p-value = 0.001 and 34 degrees of freedom; the degrees of freedom are the sum of the sample sizes minus the number of parameters, which in a two-sample comparison are two plus the number of covariates used in the original comparison).
 3. Use positive t-values for peaks of increase (e.g., activation) and negative t-values for peaks of decrease (e.g., deactivation). See **Table 2** for guidance on how to decide the sign of the t-values.
NOTE: We obtained information from studies "Heuvel" and "Soriano-Mas" after personal communication.

5. Introduction of data into SDM-PSI

1. Open **SdmPsiGui** and close the **About** splash window (avoid having any key pressed while closing it).
2. Click the **Change meta-analysis** button at the upper-left part of the graphical interface to select a directory for the meta-analysis (any new empty directory of choice will do).
3. Click the button **SDM table editor** to input general information from the studies, including their identification (column "study"), their sample sizes (columns "n1" and "n2"), the t-value that they used as statistical thresholds (column "t_thr"), and other potential variables to conduct subgroup analyses or meta-regressions.
4. Within the selected directory, create a text file for each study with the coordinates and t-value of each peak:

1. Open a text editor to create a text file named as [the identification of study] + "." + [the software] + "_" + [the stereotactic space] + ".txt". For instance, for the study "Carmona" that was conducted with SPM and reports the coordinates in MNI space, the name of the text file must be "Carmona.spm_mni.txt". If the study has no peaks, the software and stereotactic space can be replaced by "no_peaks".
2. Write the coordinates and t-value of each peak in a different row. For instance, the first rows of the text file "Carmona.spm_mni.txt" should be:
40,39,21,-5.14
53,27,21,-3.77
56,23,20,-3.63

6. Pre-processing

1. Click in the **Preprocessing** button at the left menu bar, select the modality of the studies at the list box labeled "Modality" and press **OK**. In the example meta-analysis, the modality is "VBM – gray matter".
2. Wait (some minutes) while SDM-PSI calculates the maps of the lower and upper bounds of potential effect sizes. SdmPsiGui will show four progress bars that display the status of the execution and the expected remaining time for the current process. During the calculations, the color of the circle next to "Processing status" will be yellow, and will change to green if the execution ends successfully, or to red if it fails.

7. Main analysis

1. Click the **Mean** button at the left menu bar and press **OK**.
2. Wait (some minutes) while SDM-PSI conducts the multiple imputation and meta-analysis (**Figure 1**).
3. Click the **Threshold** button in the left menu toolbox, select the uncorrected p-values of the main analysis ("MyTest_uncorrp" by default) and press **OK**. SDM-PSI will automatically open both MRICron to visualize the results and a webpage with a detailed report of them.
4. Press the **FWE correction** button at the left menu toolbox, select the main analysis in the list box ("MyTest" by default) and press **OK**.
5. Wait (some hours or even days) while SDM-PSI conducts the permutation test.
6. Click the **Threshold** button in the left menu toolbox, select the TFCE-correction of the main analysis ("MyTest_corrp_tfce" by default) and press **OK**. SDM-PSI will automatically open both MRICron to visualize the results and a webpage with a detailed report of them.

8. Heterogeneity, publication bias and grading

1. Click the **Extract** button on the left menu toolbox, select a peak from the main analysis ("MyTest_z_p_0.05000_10_neg_peak1" by default) and press **OK**. SDM-PSI will automatically open a webpage with statistics of this peak. Write down the heterogeneity I^2 statistic.
2. Click the **Bias Test** button on the left menu toolbox, select a peak from the main analysis ("MyTest_z_p_0.05000_10_neg_peak1" by default) and press **OK**. SDM-PSI will automatically open a webpage with a funnel plot and the results of a test for small-study effect and a test for excess significance. The former tests whether there is asymmetry in the funnel plot (i.e., larger effect size in small studies), which could indicate that small studies are only published if they find large effect sizes or other sources of bias. The latter tests whether the number of studies with statistically significant results is larger than expected, which could indicate that studies are only published if they find statistically significant results or other sources of bias.
3. Press the **Evidence grading** button from the top toolbox, select the main analysis ("MyTest" by default) from the list box and press **OK**. After some seconds, SDM-PSI will automatically open MRICron to show the classes of evidence.

Representative Results

As shown in the map opened in MRICron when thresholding of the main analysis (**Figure 2**, step 7.6), patients with OCD had statistically significantly smaller gray matter volume in dorsal anterior cingulate/medial frontal cortex. The accompanying webpage details that the cluster is moderately small (143 voxels) and mainly located at Brodmann area 32, and that the peak of the cluster is at MNI [2, 32, 32], has a z-value of -4.97 and a FWER-corrected p-value of 0.01.

In the webpages obtained in steps 8.1 and 8.2, the low I^2 statistic (1.5%) indicates very small heterogeneity, the funnel plot does not show asymmetries (**Figure 3**), and both the test for small-study effect and the test for excess of significance are negative. However, the evidence of smaller gray matter volume in dorsal anterior cingulate cortex is moderately weak, as shown in the map open in MRICron when grading evidence, especially due to limited amount of data.

When the main analysis was thresholded using a less stringent statistical significance level (step 7.3), patients also showed statistically significantly larger gray matter volume in the striatum and superior parietal gyrus (uncorrected p-values = 0.00006 and 0.0002 respectively), but the evidence of these abnormalities should be considered weaker.

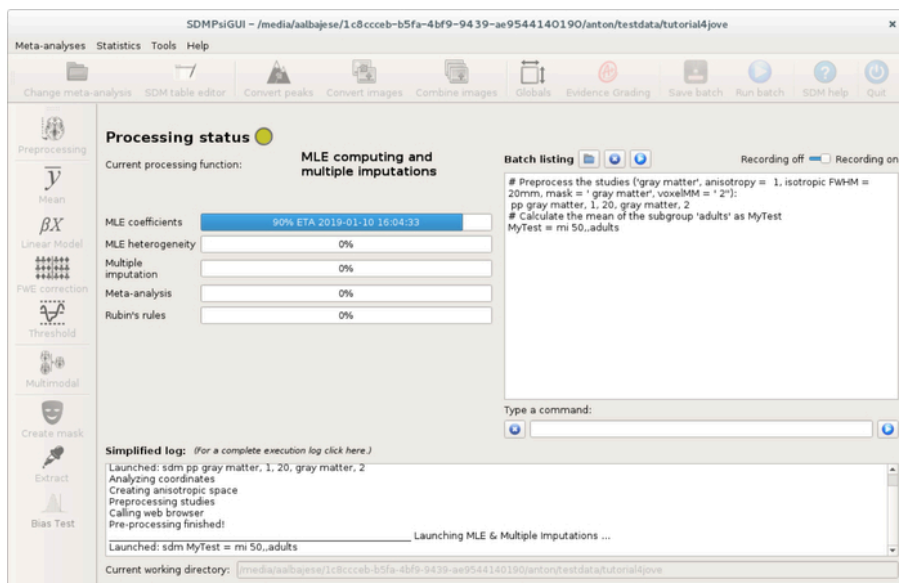


Figure 1: Main window of the SDP-PSI graphical user interface during a mean execution. [Please click here to view a larger version of this figure.](#)

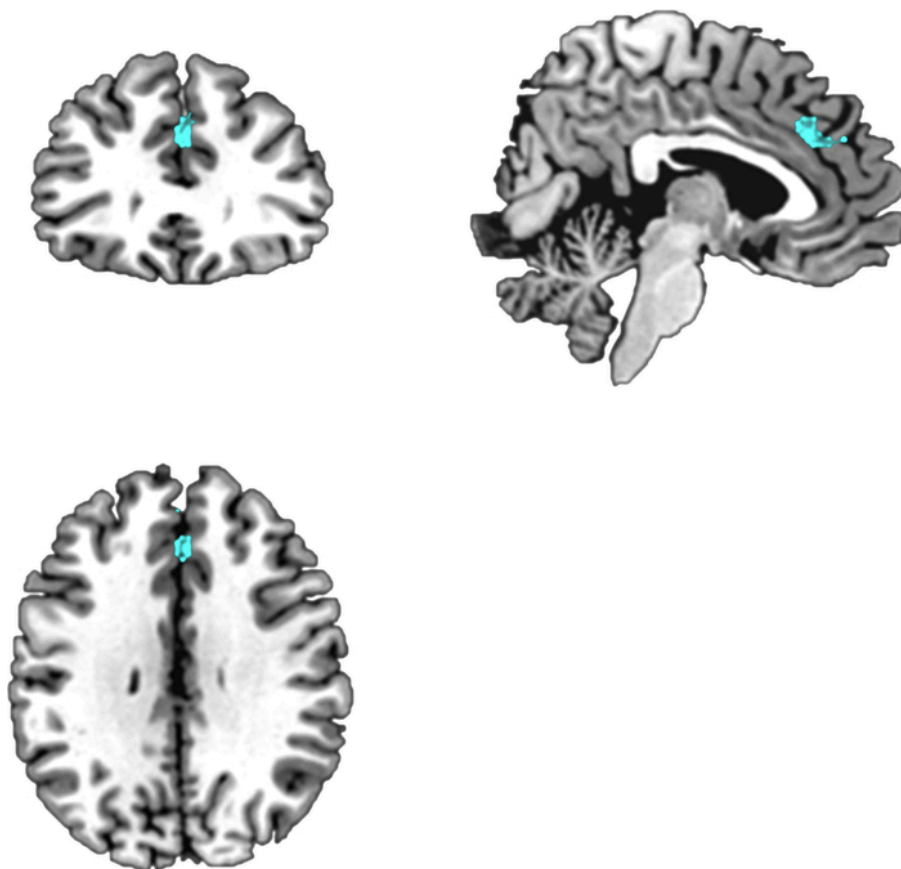


Figure 2: Regions of statistically significant smaller gray matter volume in patients with obsessive-compulsive disorder as compared to matched healthy controls.

The cluster of statistical significance covers 143 voxels, has its peak at MNI [2,32,32], and includes mostly dorsal anterior cingulate/medial frontal cortex, Brodmann area 32. [Please click here to view a larger version of this figure.](#)

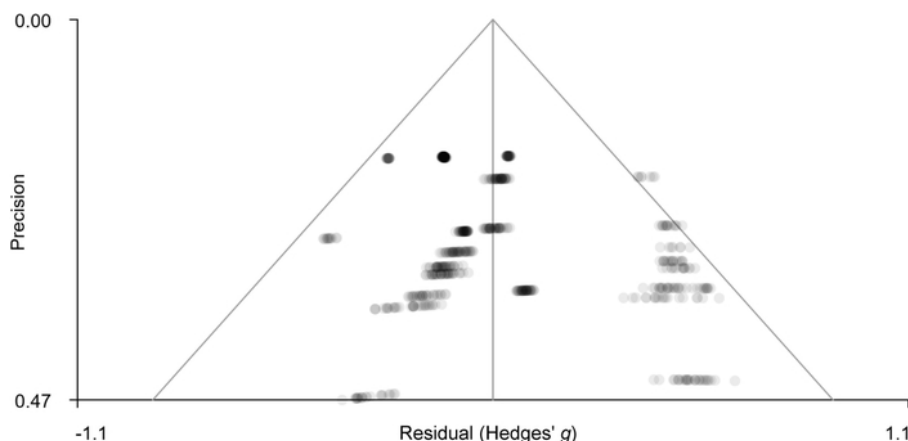


Figure 3: Funnel plot for the peak of the cluster of statistically significantly smaller gray matter volume in dorsal anterior cingulate cortex. Please click here to view a larger version of this figure.

Software packages	Coding in SDM
Statistical Parametric Mapping (SPM)	spm
FMRIB Software Library (FSL)	fsl
Other packages	other
Stereotactic space	Coding in SDM
Montreal Neurological Institute (MNI)	mni
Raw Talairach	tal
MNI converted to Talairach using Brett transform	brett

Table 1: List of software packages and stereotactic spaces understood by SDM-PSI.

	t-values must be positive when:	t-values must be negative when:
One-sample fMRI studies	task > baseline (activation)	task < baseline (deactivation)
Two-sample fMRI studies	patients > controls in task > baseline (hyper-activation)	patients < controls in task > baseline (hypo-activation)
	patients < controls in task < baseline (failure of deactivation)	patients > controls in task < baseline (hyper-deactivation)
Two-sample VBM / FA studies	patients > controls (larger volume / FA)	patients < controls (smaller volume / FA)

Table 2: Sign of the t-values of the peaks.

Discussion

As introduced earlier, most voxel-based meta-analytic methods use a test for convergence of peaks that has some limitations, and then conduct a binary classification of the evidence exclusively based on p-values.

In this protocol, we detailed how to conduct a voxel-based meta-analysis using SDM-PSI, which has a number of positive features including a standard permutation test to assess the statistical significance of the effects. In addition, we show how the strength of the evidence can be graded using a set of criteria that go beyond a binary classification that solely relies on one statistical significance level.

To facilitate the replication of the example meta-analysis, we provide the data already extracted from the manuscripts from a previous meta-analysis. Interestingly, in the manuscript of that meta-analysis, the evidence “seemed” stronger than the evidence that we found with the updated methods. We therefore suggest that unsystematic evaluations of the evidence in previous voxel-based meta-analyses are taken with caution.

We hope that following this protocol, neuroimaging meta-analyses provide a richer and more granulate description of the evidence of neuroimaging findings.

Disclosures

The authors have nothing to disclose.

Acknowledgments

This work was supported by Miguel Servet Research Contract MS14/00041 and Research Project PI14/00292 from the Plan Nacional de I+D+i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación, the European Regional Development Fund (FEDER), and by PFIS Predoctoral Contract FI16/00311. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

1. Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage*. **16** (3 Pt 1), 765-780 (2002).
2. Laird, A. R. et al. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*. **25** (1), 155-164 (2005).
3. Wager, T. D., Lindquist, M., & Kaplan, L. Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*. **2** (2), 150-158 (2007).
4. Radua, J., Mataix-Cols, D. Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *British Journal of Psychiatry*. **195** (5), 393-402 (2009).
5. Eickhoff, S. B. et al. 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).
6. Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., Fox, P. T. Activation likelihood estimation meta-analysis revisited. *Neuroimage*. **59** (3), 2349-2361 (2012).
7. Albajes-Eizagirre, A., Radua, J. What do results from coordinate-based meta-analyses tell us? *Neuroimage*. **176**, 550-553 (2018).
8. Albajes-Eizagirre, A., Solanes, A., Vieta, E., Radua, J. Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM. *Neuroimage*. **186**, 174-184 (2018).
9. Bossier, H. et al. The Influence of Study-Level Inference Models and Study Set Size on Coordinate-Based fMRI Meta-Analyses. *Frontiers in Neuroscience*. **11**, 745 (2017).
10. Radua, J. et al. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European Psychiatry*. **27** (8), 605-611 (2012).
11. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., Nichols, T. E. Permutation inference for the general linear model. *Neuroimage*. **92**, 381-397 (2014).
12. Smith, S. M., Nichols, T. E. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*. **44** (1), 83-98 (2009).
13. Radua, J. et al. Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Frontiers in Psychiatry*. **5**, 13 (2014).
14. Albajes-Eizagirre, A., Solanes, A., Radua, J. Meta-analysis of non-statistically significant unreported effects. *Statistical Methods in Medical Research*. 962280218811349 (2018).
15. Durnez, J., Moerkerke, B., Nichols, T. E. Post-hoc power estimation for topological inference in fMRI. *Neuroimage*. **84**, 45-64 (2014).
16. Fullana, M. A. et al. Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neuroscience & Biobehavioral Reviews*. **88**, 16-25 (2018).
17. Wise, T. et al. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. *Molecular Psychiatry*. **22** (10), 1455-1463 (2017).
18. Radua, J. et al. Multimodal voxel-based meta-analysis of white matter abnormalities in obsessive-compulsive disorder. *Neuropsychopharmacology*. **39** (7), 1547-1557 (2014).
19. He, W. et al. Meta-analytic comparison between PIB-PET and FDG-PET results in Alzheimer's disease and MCI. *Cell Biochemistry and Biophysics*. **71** (1), 17-26 (2015).
20. Muller, V. I. et al. Ten simple rules for neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*. **84**, 151-161 (2018).
21. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal*. **339**, b2535 (2009).
22. Radua, J. et al. What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry*. **17**, 49-66 (2018).

Chapter 4

Discussion

As exposed in the introduction chapter, the objective of this thesis was the development of an improved method for meta-analysis of neuroimaging studies that overcomes the drawbacks of previous neuroimaging meta-analytic methods. Four were the planned stages to be deployed to attain this objective and, therefore, I shall discuss the results attained in each of such stages before discussing the overall conclusions.

In our publication “What do results from coordinate-based meta-analyses tell us?” (Chapter 3 article 1) I showed that most coordinate-based meta-analysis methods in the time of conception of the thesis tested for spatial convergence of voxel activations, rather than testing for voxel not-null activation. Put in plainer words, they tested whether a voxel “activates more than most voxels do” instead of whether a voxel simply “activates”. I also showed that testing for spatial convergence may have two drawbacks because it relies on special assumptions that sometimes may not be met. All these arguments, well supported by the analyses and the practical example provided, backed up our conclusion of the need of applying standard subject-based permutation tests on our methodology.

Continuing into the second step of our working plan, we intended to improve the recreation of original maps for studies only reporting coordinate-based results. The chosen strategy was to implement a modified, improved version of the MetaNSUE methodology, which applies likelihood estimation and multiple imputations for the estimation of the missing information. In our publication “Meta-analysis of non-statistically significant unreported effects” (Chapter 3 article 2), I explained the special constraints of the application of MetaNSUE in neuroimaging meta-analyses. Briefly, in the latter, there is often a scarcity of known data, due to the usual small proportion of studies from which full statistical map results are available. Besides, there is a potential presence of extremely high t-values in the results of neuroimage studies. To overcome these specific

constraints, we conducted a substantial modification of the MetaNSUE methodology. The upgraded MetaNSUE methodology obtained in this thesis presents very satisfactory results, with negligible or even null bias on the estimation of the effect size or, also very important, the heterogeneity. The validation analysis was performed comparing in multiple simulations the results obtained meta-analyzing using all data from studies (i.e. there was no missing information) with the results obtained with MetaNSUE after removing t-values from non-statistically significant studies. The results were consistent across simulations with varying effect sizes and heterogeneities, showing the robustness and power of the obtained methodology.

Advancing through our timeline, we reached the paramount objective of this thesis, which was the development and implementation of a new method for the meta-analysis of neuroimage studies that overcomes the drawbacks of previous methods. The development of the methodology involved statistical research covered in the first and second stage described above, whereas the actual software implementation involved on top of those aspects of algorithms and of programming that required joining the statistical knowledge to the computer science expertise to find a viable design and implementation of the methodology. On the one hand, the application of MetaNSUE for neuroimaging data raises the requirement of combining the accuracy of the imputations with a realistic local spatial structure. To attain such a challenging combination, a solution inspired in Gaussian random fields was designed and successfully validated by our simulations. The accurate implementation of MetaNSUE and the resulting accurate recreation of original study maps allowed us to develop an algorithm to perform the imputation of individual subject images for each study. Such imputation was designed to respect the expected correlation between adjacent voxels according to our correlation template (Radua et al. (2014)). As explained in our publication “Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM”, the algorithm needed to be designed taking into consideration computational performance to be able to obtain a viable implementation in terms of execution requirements (see article 3 in Chapter 3 for details).

On the other hand, once we had a methodology providing the imputation of individual subject images we could develop the application of a standard permutation test to perform the family-wise error (FWE) correction to obtain the p-value significance map of the meta-analysis results. Furthermore, we implemented the FWE for both voxel-wise correction and threshold-free cluster enhancement (TFCE) correction. The implementation of such demanding algorithms as the standard permutation tests also required computer engineering work to attain a scalable and time-wise viable application. Our methodology aimed at the general community of neuroimaging meta-analysts and, therefore, needed to be executed in computers commonly available to researchers rather than

in high-performance computing clusters. A highly optimized multi-threading implementation was developed and coded from scratch, enabling the execution of thousands of permutations tests in conventional customer-level desktop computers.

The final validation of our methodology and our application of the standard permutation test yielded very satisfactory results. We tested for different scenarios involving different study sizes and meta-analysis sizes, and the results show that the new methodology controls the FWE rate below 5% in all the scenarios we tested for voxel-wise and for TFCE corrections. This means that in the absence of true effects, an SDM-PSI meta-analysis should rarely detect false effects. It is well worth mentioning that we proposed the new algorithm as a general framework and published it as such, giving its implementation with SDM as an example of a viable derived methodology. As already introduced in the article, the permutations of the subject images (PSI) algorithm could be implemented to several current CBMA such as ALE or MKDA.

The ultimate purpose of developing a new methodology is to provide all its potential to the neuroimaging meta-analysts. One indispensable condition for a method to become useful and powerful for the meta-analysts is, indeed, to be used. And for a method or software to be used it, is needed that the researchers know about it. And not only know about it but know how to use it. Therefore, we also dedicated great efforts during this thesis to develop a user-friendly graphical interface for the software, to record and publish a visual article describing practically how to use the software (Chapter 3 article 4), and to provide user support through the forum for the the project and other media. We understood that developing a good methodology was a big part of the job but giving the software life and making it easy for users to use it was also part of it. At the time of finishing this thesis, there are already several published studies that used our implementation of the SDM-PSI method (Suchting et al. (2020), Kunimatsu, Yasaka, Akai, Kunimatsu, and Abe (2019), Enge, Friederici, and Skeide (2020)), and many more in print. SDM-PSI has already been used for research areas like drug use disorders (Suchting et al. (2020)), post-traumatic disorders (Kunimatsu et al. (2019)), language comprehension (Enge et al. (2020)), fear conditioning, cortical gyrification on autism and attention deficit hyperactivity disorder (ADHD), psychosis, and many others. The modalities of meta-analyses already performed by SDM-PSI also cover almost any compatible one, like fMRI, DTI, VBM, cortical thickness, etc.

There are some limitations of the SDM-PSI implementation presented in this thesis that are worth discussing at this point. First, it requires to know the t-threshold used by the authors of the studies. For those cases in which the t-threshold is not known, there is the conservative recommendation of choosing the t-value corresponding to $p=0.001$ uncorrected. Second, SDM-PSI does not account for statistically significant results that

were not reported due to small cluster extent (they did not form a cluster large enough for the original authors to report it). However, we estimate the impact of this issue on the meta-analysis to be limited and conservative. To our knowledge, no other methodology accounts for such non-reported results either. And third, some adjustments were made on MetaNSUE to avoid single studies to have an erroneously large influence on the meta-analysis in some scenarios. Such adjustments make the new method slightly bias downwards the effect sizes and z-values. However, we believe that the prevention of a single or a few studies having an erroneously large influence is more important than the resulting minor conservative bias. Even more, considering that this conservative bias of the effect sizes and z-values does not affect the FWE correction.

In summary, I can honestly say that I am confident that this thesis achieved its objectives and that the obtained methodology solves important open issues in the meta-analyses of neuroimaging studies. Of importance should be the improvements offered in terms of dealing with missing information, an issue of high impact in the area, and the improvement in terms of accuracy of the estimation of the significance of the results, another sensible problem. On the whole, SDM-PSI highly improves meta-analysis combining coordinate-based and full statistical maps studies, accurately estimates the statistical significance maps of the results, improves the tools to detect possible bias already present in previous versions of SDM, improves the capabilities of previous versions to perform meta-regression, offering full flexibility to specify the regression models, and provides the framework to further developments. Despite the finiteness of this thesis, the work on the upgrading and extension of the methodology does not end with it. It is a work in progress and it includes the development of further tests to detect more of the bias discussed in this thesis, like the excessive significance test (see Chapter 1), the development of additional and improved brain templates, the extension of linear models offered by the software (like linear models without intercept), and many other new features and improvements that the several open projects within the global SDM project may bring.

Chapter 5

Conclusions

1. Most used coordinate-based meta-analysis methods test for spatial convergence rather than for non-null activation of voxels. This biases the results.
2. MetaNSUE attains, by using likelihood estimation and multiple imputations, good results in problematic scenarios common to neuroimaging studies: the scarcity of known data and the potential presence of high t-values. Validation results confirmed negligible or null bias on the estimation of effect size and heterogeneity.
3. We provide a viable and efficient implementation of the MetaNSUE algorithm applicable to neuroimaging data.
4. Our implementation of the MetaNSUE method also provides the tools needed to enable the conduction of a standard permutation test.
5. There exists an algorithm (PSI) that enables an efficient implementation of the standard permutation test on neuroimaging data.
6. The validations prove that our method controls the FWE rate well below 5% for both voxel-wise and TFCE.
7. The proposed PSI algorithm can also be applied to other current CBMA methods, such as ALE or MKDA.
8. SDM-PSI has already been used in ([Suchting et al. \(2020\)](#), [Kunimatsu et al. \(2019\)](#), [Enge et al. \(2020\)](#)) published papers.

References

- Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., . . . Penuliar, M. (2015). Estimating the Reproducibility of Psychological Science. *Science*, *349*. doi: 10.1126/science.aac4716
- Acar, F., Seurinck, R., Eickhoff, S. B., & Moerkerke, B. (2018, nov). Assessing robustness against potential publication bias in Activation Likelihood Estimation (ALE) meta-analyses for fMRI. *PLoS ONE*, *13*(11). doi: 10.1371/journal.pone.0208177
- Aitkenhead, D. (2013). Peter Higgs: I wouldn't be productive enough for today's academic system. *The Guardian*, 6–7. Retrieved from <https://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system><http://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system/print>
- AJG “The Negative Issue” November 2016 - American College of Gastroenterology. (n.d.). Retrieved 2020-04-22, from <https://gi.org/2016/11/01/the-american-journal-of-gastroenterology-presents-the-negative-issue/>
- Al et, S. J. A. (2016). *Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials.* - PubMed - NCBI. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21784880>
- Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992, jul). A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction. *JAMA: The Journal of the American Medical Association*, *268*(2), 240–248. doi: 10.1001/jama.1992.03490020088036
- ASA Develops Reproducible Research Recommendations. (n.d.). Retrieved 2020-04-28, from <https://www.amstat.org/ASA/News/ASA-Develops-Reproducible-Research-Recommendations.aspx>
- Ashby, F. G. (2019). *Statistical analysis of fMRI data.*
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–454.
- Bearden, C. E., & Thompson, P. M. (2017). Emerging Global Initiatives in Neurogenetics: The Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA) Consortium. *Neuron*, *94*(2), 232–236. Retrieved from <http://www>

- [.sciencedirect.com/science/article/pii/S0896627317302453](https://www.sciencedirect.com/science/article/pii/S0896627317302453) doi: <https://doi.org/10.1016/j.neuron.2017.03.033>
- Begg, C. B., & 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. Retrieved from <https://www.jstor.org/stable/2982993> doi: 10.2307/2982993
- Begley, C. G., & Ellis, L. M. (2012, mar). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. Retrieved from <https://doi.org/10.1038/483531a> doi: 10.1038/483531a
- Beisteiner, R., Pernet, C., & Stippich, C. (2019). Can We Standardize Clinical Functional Neuroimaging Procedures? *Frontiers in Neurology*, *9*, 1153. Retrieved from <https://www.frontiersin.org/article/10.3389/fneur.2018.01153> doi: 10.3389/fneur.2018.01153
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.
- Bennett, C. M., & Miller, M. B. (2010). *How reliable are the results from functional magnetic resonance imaging?* (Vol. 1191). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20392279> doi: 10.1111/j.1749-6632.2010.05446.x
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley and Sons. Retrieved from <http://doi.wiley.com/10.1002/9780470743386> doi: 10.1002/9780470743386
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... others (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 1–7.
- brainmap.org — Home*. (n.d.). Retrieved 2020-04-22, from <http://www.brainmap.org/>
- Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014, may). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. Retrieved from <http://journals.sagepub.com/doi/10.1177/1745691614529796> doi: 10.1177/1745691614529796
- Carp, J. (2012). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, *6*, 149. Retrieved from <https://www.frontiersin.org/article/10.3389/fnins.2012.00149> doi: 10.3389/fnins.2012.00149
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.
- Collins, F. S., & Tabak, L. A. (2014, jan). *NIH plans to enhance reproducibility* (Vol. 505) (No. 7485). doi: 10.1038/505612a

- Costafreda, S. G., David, A. S., & Brammer, M. J. (2009). A parametric approach to voxel-based meta-analysis. *NeuroImage*, *46*(1), 115–122. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18985131> doi: 10.1016/j.neuroimage.2009.01.031
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
- David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., ... Ioannidis, J. P. (2013, jul). Potential Reporting Bias in fMRI Studies of the Brain. *PLoS ONE*, *8*(7), e70104. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23936149><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3723634> doi: 10.1371/journal.pone.0070104
- Decullier, E., Lhéritier, V., & Chapuis, F. (2005, jul). Fate of biomedical research protocols and publication bias in France: Retrospective cohort study. *British Medical Journal*, *331*(7507), 19–22. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15967761><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC558532> doi: 10.1136/bmj.38488.385995.8F
- Dickersin, K., & Min, Y. I. (1993). NIH clinical trials and publication bias. *The Online journal of current clinical trials*, *Doc No 50*. Retrieved from <https://www.researchgate.net/publication/14893304-NIH-clinical-trials-and-publication-bias>
- Dirnagl, U., & Lauritzen, M. (2010, jul). *Editorial: Fighting publication bias: Introducing the Negative Results section* (Vol. 30) (No. 7). doi: 10.1038/jcbfm.2010.51
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple , graphical test measures of funnel plot asymmetry. *Bmj*, *315*(7109), 629–34. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9310563><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2127453> doi: 10.1136/bmj.315.7109.629
- Eickhoff, S., Nichols, T. E., Van Horn, J. D., & Turner, J. A. (2016, jan). *Sharing the wealth: Neuroimaging data repositories* (Vol. 124). Academic Press Inc. doi: 10.1016/j.neuroimage.2015.10.079
- Enge, A., Friederici, A. D., & Skeide, M. A. (2020). A meta-analysis of fmri studies of language comprehension in children. *NeuroImage*, 116858.
- Eysenck, H. J. (1952). The effects of psychotherapy: an evaluation. *Journal of consulting psychology*, *16*(5), 319.
- Eysenck, H. J. (1994, sep). *Meta-analysis and its problems* (Vol. 309) (No. 6957). British Medical Journal Publishing Group. doi: 10.1136/bmj.309.6957.789

- Filler, A. (2009). The History, Development and Impact of Computed Imaging in Neurological Diagnosis and Neurosurgery: CT, MRI, and DTI. *Nature Precedings*. Retrieved from <http://precedings.nature.com/documents/3267/version/5> doi: 10.1038/npre.2009.3267.5
- Fostering reproducible fmri research. (2017, mar). *Nature Neuroscience*, 20(3), 298–298. Retrieved from <https://doi.org/10.1038/nn.4521><http://www.nature.com/articles/nn.4521> doi: 10.1038/nn.4521
- General Information — eNeuro. (n.d.). Retrieved 2020-04-22, from <https://www.eneuro.org/content/general-information{#}types>
- Ghiasi, G., Larivière, V., & Sugimoto, C. R. (2015, dec). On the compliance of women engineers with a gendered scientific system. *PLoS ONE*, 10(12), e0145931. Retrieved from <https://dx.plos.org/10.1371/journal.pone.0145931> doi: 10.1371/journal.pone.0145931
- Gilbert, R., Salanti, G., Harden, M., & See, S. (2005). *Infant sleeping position and the sudden infant death syndrome: Systematic review of observational studies and historical review of recommendations from 1940 to 2002* (Vol. 34) (No. 4). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15843394> doi: 10.1093/ije/dyi088
- Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L., Auerbach, E. J., Behrens, T. E., ... Van Essen, D. C. (2016, sep). *The Human Connectome Project's neuroimaging approach* (Vol. 19) (No. 9). Nature Publishing Group. doi: 10.1038/nn.4361
- Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., ... Margulies, D. S. (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9, 8. Retrieved from <https://www.frontiersin.org/article/10.3389/fninf.2015.00008> doi: 10.3389/fninf.2015.00008
- Guyatt, G. H., Oxman, A. D., Montori, V., Vist, G., Kunz, R., Brozek, J., ... Schünemann, H. J. (2011, dec). GRADE guidelines: 5. Rating the quality of evidence - Publication bias. *Journal of Clinical Epidemiology*, 64(12), 1277–1282. doi: 10.1016/j.jclinepi.2011.01.011
- Guyatt, G. H., Oxman, A. D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., ... Schünemann, H. J. (2011). GRADE guidelines: 4. Rating the quality of evidence - Study limitations (risk of bias). *Journal of Clinical Epidemiology*, 64(4), 407–415. doi: 10.1016/j.jclinepi.2010.07.017
- Harris, J. K., Combs, T. B., Johnson, K. J., Carothers, B. J., Luke, D. A., & Wang, X. (2019). *Three Changes Public Health Scientists Can Make to Help Build a Culture of Reproducible Research* (Vol. 134) (No. 2). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/30657732> doi: 10.1177/0033354918821076

- Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., & Wang, X. (2018, sep). Use of reproducible research practices in public health: A survey of public health analysts. *{PLOS} {ONE}*, *13*(9), e0202447. Retrieved from <https://doi.org/10.1371/journal.pone.0202447> doi: 10.1371/journal.pone.0202447
- Hart, B., Lundh, A., & Bero, L. (2012, jan). Effect of reporting bias on meta-analyses of drug trials: Reanalysis of meta-analyses. *BMJ (Online)*, *344*(7838). doi: 10.1136/bmj.d7202
- Hedges, L. V., & Schauer, J. M. (2019, oct). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. Retrieved from <https://doi.org/10.1037/met0000189> doi: 10.1037/met0000189
- Hopewell, S., Clarke, M. J., Stewart, L., & Tierney, J. (2007). *Time to publication for results of clinical trials* (No. 2). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17443632> doi: 10.1002/14651858.MR000011.pub2
- Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). *Publication bias in clinical trials due to statistical significance or direction of trial results* (No. 1). John Wiley and Sons Ltd. doi: 10.1002/14651858.MR000006.pub3
- Improving Reproducibility and Replicability - Reproducibility and Replicability in Science - NCBI Bookshelf*. (n.d.). Retrieved 2020-04-28, from <https://www.ncbi.nlm.nih.gov/books/NBK547525/>
- Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). *Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention* (Vol. 18) (No. 5). Elsevier Ltd. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24656991> doi: 10.1016/j.tics.2014.02.010
- Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15878467> doi: 10.1016/j.jclinepi.2004.10.019
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. doi: 10.1177/1740774507079441
- Ioannidis, J. P. A. (2005, aug). Why most published research findings are false. *PLoS medicine*, *2*(8), e124–e124. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/16060722> doi: 10.1371/journal.pmed.0020124
- IRIS Replication Award- IRIS Digital Repository*. (n.d.). Retrieved 2020-04-22, from [https://www.iris-database.org/iris/app/home/replication\[_\]award](https://www.iris-database.org/iris/app/home/replication[_]award)
- Irwig, L., Macaskill, P., Berry, G., & Glasziou, P. (1998). *Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased*. (Vol. 316) (No. 7129). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9492687>

- J. Light, R., & B. Pillemer, D. (1986, oct). Summing Up: The Science of Reviewing Research Harvard University Press: Cambridge, MA, 1984, xiii+191 pp. *Educational Researcher*, 15(8), 16–17. doi: 10.3102/0013189x015008016
- Jefferson, L., Fairhurst, C., Cooper, E., Hewitt, C., Torgerson, T., Cook, L., ... Torgerson, D. (2016, oct). No difference found in time to publication by statistical significance of trial results: a methodological review. *JRSM Open*, 7(10), 205427041664928. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/27757242><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5052771> doi: 10.1177/2054270416649283
- Jennings, R. G., & Van Horn, J. D. (2012). Publication bias in neuroimaging research: Implications for meta-analyses. *Neuroinformatics*, 10(1), 67–80. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23936149><https://www.ncbi.nlm.nih.gov/pubmed/21643733> doi: 10.1007/s12021-011-9125-y
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. Retrieved from <https://doi.org/10.1177/0956797611430953> doi: 10.1177/0956797611430953
- Johnson, G. (2014, jan). New Truths That Only One Can See. *The new york times*. Retrieved from <https://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html>
- Kang, J., Johnson, T. D., Nichols, T. E., & Wager, T. D. (2011, mar). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *Journal of the American Statistical Association*, 106(493), 124–134. doi: 10.1198/jasa.2011.ap09735
- Kang, J., Nichols, T. E., Wager, T. D., & Johnson, T. D. (2014). A bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *Annals of Applied Statistics*, 8(3), 1561–1582. doi: 10.1214/14-AOAS757
- Kennedy, D. N. (2018, apr). *Neuroimaging neuroinformatics: Sample size and other evolutionary topics* (Vol. 16) (No. 2). Humana Press Inc. doi: 10.1007/s12021-018-9379-8
- Kicinski, M. (2014). How does under-reporting of negative and inconclusive results affect the false-positive rate in meta-analysis? A simulation study. *BMJ Open*, 4(8). doi: 10.1136/bmjopen-2014-004831
- Kicinski, M., Springate, D. A., & Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, 34(20), 2781–2793. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6525> doi: 10.1002/sim.6525
- Kirchherr, J. (2017, jun). Why we can't trust academic journals to tell the scientific truth. *The guardian*. Retrieved from <https://www.theguardian.com/>

higher-education-network/2017/jun/06/why-we-cant-trust-academic-journals-to-tell-the-scientific-truth

- Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010, mar). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ (Online)*, *340*(7747), 637–640. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20156912> doi: 10.1136/bmj.c365
- Kunimatsu, A., Yasaka, K., Akai, H., Kunimatsu, N., & Abe, O. (2019). Mri findings in posttraumatic stress disorder. *Journal of Magnetic Resonance Imaging*.
- Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992, jul). Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction. *New England Journal of Medicine*, *327*(4), 248–254. Retrieved from <http://www.nejm.org/doi/abs/10.1056/NEJM199207233270406> doi: 10.1056/NEJM199207233270406
- Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *Bmj*, *333*(7568), 597–600.
- Leeds, N. E., & Kieffer, S. A. (2000). Evolution of diagnostic neuroradiology from 1904 to 1999. *Radiology*, *217*(2), 309–318. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/11058623> doi: 10.1148/radiology.217.2.r00nv45309
- Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics*, *74*(3), 785–794. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29141096> doi: 10.1111/biom.12817
- Loannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association*, *279*(4), 281–286. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9450711> doi: 10.1001/jama.279.4.281
- Lorca-Puls, D. L., Gajardo-Vidal, A., White, J., Seghier, M. L., Leff, A. P., Green, D. W., ... Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. *Neuropsychologia*, *115*, 101–111. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29550526> doi: 10.1016/j.neuropsychologia.2018.03.014
- The Missing Pieces: A Collection of Negative, Null and Inconclusive Results*. (n.d.). Retrieved 2020-04-22, from <https://collections.plos.org/missing-pieces>
- Montagna, S., Wager, T., Barrett, L. F., Johnson, T. D., & Nichols, T. E. (2018, mar). Spatial Bayesian latent factor regression modeling of coordinate-based meta-analysis data. *Biometrics*, *74*(1), 342–353. doi: 10.1111/biom.12713
- Mueller, K. F., Meerpohl, J. J., Briel, M., Antes, G., von Elm, E., Lang, B., ... Bassler, D. (2013, jul). Detecting, quantifying and adjusting for publication bias in meta-analyses: protocol of a systematic review on methods. *Systematic reviews*, *2*,

60. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23885765><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3733739> doi: 10.1186/2046-4053-2-60
- Mumford_Keep Calm and Scan On - organization for human brain mapping.pdf.* (n.d.). Retrieved from <https://www.ohbmbrianmappingblog.com/blog/keep-calm-and-scan-on>
- Munafò, M., Noble, S., Browne, W. J., Brunner, D., Button, K., Ferreira, J., ... Blumenstein, R. (2014, sep). *Scientific rigor and the art of motorcycle maintenance* (Vol. 32) (No. 9). Nature Publishing Group. doi: 10.1038/nbt.3004
- Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, 2(1), 130. Retrieved from <https://doi.org/10.1038/s42003-019-0378-6> doi: 10.1038/s42003-019-0378-6
- Negative Results — Home page of Scientific Journal Negative Results.* (n.d.). Retrieved 2020-04-22, from <https://www.negative-results.org/>
- Niven, D. J., McCormick, T. J., Straus, S. E., Hemmelgarn, B. R., Jeffs, L., Barnes, T. R. M., & Stelfox, H. T. (2018, feb). Reproducibility of clinical research in critical care: a scoping review. *{BMC} Medicine*, 16(1). Retrieved from <https://doi.org/10.1186/s12916-018-1018-6> doi: 10.1186/s12916-018-1018-6
- Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6), 528–530. Retrieved from <https://doi.org/10.1177/1745691612465253> doi: 10.1177/1745691612465253
- Perrin, S. (2014, mar). *Make mouse studies work* (Vol. 507) (No. 7493). Nature Publishing Group. doi: 10.1038/507423a
- Pfeiffer, T., Bertram, L., & Ioannidis, J. P. (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLoS ONE*, 6(3). doi: 10.1371/journal.pone.0018362
- Picciotto, M. (2018, apr). *Analytical transparency and reproducibility in human neuroimaging studies* (Vol. 38) (No. 14). Society for Neuroscience. doi: 10.1523/JNEUROSCI.0424-18.2018
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126. Retrieved from <https://doi.org/10.1038/nrn.2016.167> doi: 10.1038/nrn.2016.167
- Popper, K. R. (1934). *The Logic of Scientific Discovery*. London: Hutchinson.
- Publication bias - Wikipedia.* (n.d.). Retrieved 2020-04-22, from <https://en.wikipedia.org/wiki/Publication{ }bias>

- Qunaj, L., Jain, R. H., Atoria, C. L., Gennarelli, R. L., Miller, J. E., & Bach, P. B. (2018). Delays in the publication of important clinical trial findings in oncology. *JAMA oncology*, *4*(7), e180264. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/29710325><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6145729> doi: 10.1001/jamaoncol.2018.0264
- Radua, J., Rubia, K., Canales-Rodríguez, E. J., Pomarol-Clotet, E., Fusar-Poli, P., & Mataix-Cols, D. (2014). Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Frontiers in Psychiatry*, *5*(FEB). doi: 10.3389/fpsy.2014.00013
- Repko, A. F., & Szostak, R. (2020). *Interdisciplinary research : process and theory. Replication Award - www.humanbrainmapping.org.* (n.d.). Retrieved 2020-04-22, from <https://www.humanbrainmapping.org/m/pages.cfm?pageid=3731>
- Rosenthal, R. (1979, may). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi: 10.1037/0033-2909.86.3.638
- Samartsidis, P., Montagna, S., Johnson, T. D., & Nichols, T. E. (2017, nov). The coordinate-based meta-analysis of neuroimaging data. *Statistical Science*, *32*(4), 580–599. doi: 10.1214/17-STS624
- Schimmack, U. (2012, dec). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. doi: 10.1037/a0029487
- Science/AAAS — Special Section: Data Replication and Reproducibility.* (n.d.). Retrieved 2020-04-22, from <https://www.sciencemag.org/site/special/data-rep><http://www.sciencemag.org/site/special/data-rep/>
- Shapin, S. (1989). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life.* Princeton University Press. Retrieved from <https://www.xarg.org/ref/a/0691024324/>
- Shuter, B., Yeh, I. B., Graham, S., Au, C., & Wang, S. C. (2008). Reproducibility of brain tissue volumes in longitudinal studies: Effects of changes in signal-to-noise ratio and scanner software. *NeuroImage*, *41*(2), 371–379. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1053811908001304> doi: 10.1016/j.neuroimage.2008.02.003
- Smith, M. L., & Glass, G. V. (1977, sep). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/921048> doi: 10.1037/0003-066X.32.9.752
- Smith, S. M., & Nichols, T. E. (2018, jan). Statistical challenges in big data human neuroimaging. *Neuron*, *97*(2), 263–268. Retrieved from <https://doi.org/10.1016/j.neuron.2017.12.018> doi: 10.1016/j.neuron.2017.12.018

- Song, F., Parekh-Bhurke, S., Hooper, L., Loke, Y. K., Ryder, J. J., Sutton, A. J., ... Harvey, I. (2009). Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies. *BMC Medical Research Methodology*, *9*(1), 79. doi: 10.1186/1471-2288-9-79
- Specht, K. (2020). *Current Challenges in Translational and Clinical fMRI and Future Directions* (Vol. 10). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/31969840> doi: 10.3389/fpsy.2019.00924
- Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0895435601003778> doi: 10.1016/S0895-4356(01)00377-8
- Stuck, A. E., Rubenstein, L. Z., & Wieland, D. (1998). *Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity.* (Vol. 316) (No. 7129).
- Suchting, R., Beard, C. L., Schmitz, J. M., Soder, H. E., Yoon, J. H., Hasan, K. M., ... Lane, S. D. (2020). A meta-analysis of tract-based spatial statistics studies examining white matter integrity in cocaine use disorder. *Addiction Biology*, e12902.
- Suñé, P., Suñé, J. M., & Montoro, J. B. (2013, jan). Positive Outcomes Influence the Rate and Time to Publication, but Not the Impact Factor of Publications of Clinical Trial Results. *PLoS ONE*, *8*(1). doi: 10.1371/journal.pone.0054583
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), 1–18. Retrieved from <https://doi.org/10.1371/journal.pbio.2000797> doi: 10.1371/journal.pbio.2000797
- Szucs, D., & Ioannidis, J. P. A. (2019). Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. *bioRxiv*. Retrieved from <https://www.biorxiv.org/content/early/2019/10/17/809715> doi: 10.1101/809715
- Terrin, N., Schmid, C. H., & Lau, J. (2005, sep). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, *58*(9), 894–901. doi: 10.1016/j.jclinepi.2005.01.006
- Toga, A. W. (2002). Neuroimage databases: The good, the bad and the ugly. *Nature Reviews Neuroscience*, *3*(4), 302–309. doi: 10.1038/nrn782
- Tsujimoto, Y., Tsutsumi, Y., Kataoka, Y., Tsujimoto, H., Yamamoto, Y., Papola, D., ... Furukawa, T. A. (2017, oct). *Association between statistical significance and time to publication among systematic reviews: A study protocol for a meta-epidemiological investigation* (Vol. 7) (No. 10). BMJ Publishing Group. doi: 10.1136/bmjopen-2017-018856

- Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., & Fox, P. (2012, jan). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, *33*(1), 1–13. doi: 10.1002/hbm.21186
- Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), 62. Retrieved from <https://doi.org/10.1038/s42003-018-0073-z> doi: 10.1038/s42003-018-0073-z
- Turner, J., Eickhoff, S., & Nichols, T. (Eds.). (2017). *Sharing the wealth: Neuroimaging data repositories, Part II* (Vol. 144) (No. B). Retrieved from <https://www.sciencedirect.com/journal/neuroimage/vol/144/part/PB>
- Virtual Special Issue On Replication Studies*. (n.d.). Retrieved 2020-04-22, from <https://www.elsevier.com/social-sciences-and-humanities/business-management-and-accounting/journals/virtual-special-issue-on-replication-studies>
- Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: Current and future directions. *Social Cognitive and Affective Neuroscience*, *2*(2), 150–158. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2555451/> doi: 10.1093/scan/nsm015
- Young, C., & Horton, R. (2005, jul). Putting clinical trials into context. *The Lancet*, *366*(9480), 107–108. Retrieved from [https://doi.org/10.1016/s0140-6736\(05\)66846-8](https://doi.org/10.1016/s0140-6736(05)66846-8) doi: 10.1016/s0140-6736(05)66846-8
- Yue, Y. R., Lindquist, M. A., & Loh, J. M. (2012, jun). Meta-analysis of functional neuroimaging data using bayesian nonparametric binary regression. *Annals of Applied Statistics*, *6*(2), 697–718. Retrieved from <http://arxiv.org/abs/1206.6674><http://dx.doi.org/10.1214/11-AOAS523> doi: 10.1214/11-AOAS523