scientific reports

Check for updates

OPEN

Advancing image-based meta-analysis through systematic use of crowdsourced NeuroVault data

Julio A. Peraza[1]✉, James D. Kent[2], Ross W. Blair[3], Jean-Baptiste Poline[4], Thomas E. Nichols[5], Alejandro de la Vega[2] & Angela R. Laird[1]✉

Image-based meta-analysis (IBMA) is a powerful method for synthesizing results from various fMRI studies. However, challenges related to data accessibility and the lack of available tools and methods have limited its widespread use. This study examined the current state of the NeuroVault repository and developed a comprehensive framework for selecting and analyzing neuroimaging statistical maps within it. By systematically assessing the quality of NeuroVault's data and implementing novel selection and meta-analysis techniques, we demonstrated the repository's potential for IBMA. We created a multi-stage selection framework that includes preliminary, heuristic, and manual image selection. We conducted meta-analyses for three distinct domains: working memory, motor, and emotion processing. The results from the three manual IBMA approaches closely resembled reference maps from the Human Connectome Project. Importantly, we found that while manual selection provides the most precise results, heuristic methods can serve as robust alternatives, especially for domains that include a heterogeneous set of fMRI tasks and contrasts, such as emotion processing. Additionally, we evaluated five different meta-analytic estimator methods to assess their effectiveness in handling spurious images. For domains characterized by heterogeneous tasks, employing a robust estimator (e.g., median) is essential. This study is the first to present a systematic approach for implementing IBMA using the NeuroVault repository. We introduce an accessible and reproducible methodology that allows researchers to make the most of NeuroVault's extensive neuroimaging resources, potentially fostering greater interest in data sharing and future meta-analyses utilizing NeuroVault data.

Keywords fMRI, IBMA, NeuroVault, Meta-analysis, Image-based meta-analysis

A key challenge in neuroimaging meta-analysis, as the scientific literature continues to grow rapidly each year, is how findings are aggregated[1]. Image-based meta-analysis (IBMA) is considered the gold standard for aggregating neuroimaging results because it combines whole-brain statistical maps to identify consistent effects across studies with greater sensitivity than the popular coordinate-based meta-analysis (CBMA)[2,3]. Despite its clear benefits, the use of IBMA has been limited by a lack of accessible data and standardized analysis frameworks. While the NeuroVault repository[4] was created to improve data availability for meta-analyses, it remains underused, as only one published IBMA has utilized NeuroVault data[5]. Several issues, including non-standardized data sharing practices, the risk of spurious images in NeuroVault[6], and the absence of appropriate tools and methods, have created a significant gap between NeuroVault's potential and its actual use for meta-analysis. Here, we introduce a comprehensive framework that closes this gap by providing systematic methods for data curation, quality control, and robust meta-analysis, making NeuroVault's extensive resources accessible for IBMA.

Neuroimaging meta-analysis encompasses several approaches for aggregating neuroimaging data[7], each suited to different levels of data availability. At the highest level, mega-analysis combines individual participant data across studies, allowing for unified statistical modeling with optimal power[8–12]. When researchers have access to subject-level time series data from multiple studies, they can perform a hierarchical three-level IBMA mega-analysis[3]. This approach can produce both fixed-effects and mixed-effects inferences at the third level, accounting for both within- and between-study variance. However, individual participant data is rarely available

[1]Department of Physics, Florida International University, Miami, FL, USA. [2]Department of Psychology, University of Texas at Austin, Austin, TX, USA. [3]Department of Psychology, Stanford University, Stanford, CA, USA. [4]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada. [5]Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, Big Data Institute, University of Oxford, Oxford, UK. ✉email: jperaza@fiu.edu; alaird@fiu.edu

due to privacy concerns, storage limitations, and data sharing restrictions[13–15]. More commonly, researchers perform traditional IBMA using group-level statistical maps from each study[2]. This approach still preserves the spatial richness of whole-brain data while working with summary statistics. Various aggregation methods can be applied, including simple averaging, Fisher's[16], or Stouffer's combination methods[17]. When even group-level statistical maps are unavailable and researchers only have access to reported peak coordinates from publications, they must resort to CBMA[18–21], where methods like activation likelihood estimation (ALE)[22], kernel density analysis (KDA)[23,24], and multilevel KDA (MKDA)[25] are commonly used to combine activation coordinates across studies.

IBMA outperforms CBMA in several aspects. For instance, CBMA works with activation foci producing information loss, and it relies on kernel-based methods to infer the model activation maps from peak coordinates, requiring spatial assumptions that fail to reproduce the actual statistical map[25]. Although a combination of kernel parameters and thresholds for CBMA could be chosen to maximize the similarity with IBMA maps, such optimal settings depend on the dataset and specific considerations from individual studies (e.g., statistical thresholds) and thus are not generalizable across meta-analyses[3]. Further, considering how underpowered most neuroimaging studies are, brain activations that fail to pass certain thresholds of significance are generally discarded in a CBMA. In contrast, IBMA methods use whole-brain statistics; thus, all existing voxel-wise statistical methods are available to analyze subject-level data within studies[2]. IBMA is known to produce richer and more detailed results, with additional brain structures that are often absent from CBMA results. IBMA also has greater power; thus, one could potentially achieve similar or even better results with a small fraction of the studies generally required in CBMA. In addition, when both the parameters and variance estimates are available, hierarchical mixed-effect models can be used to account for both within- and between-study variance[3]. Despite the clear superiority of IBMA, CBMA remains the most popular approach among the neuroimaging research community for practical reasons.

In the last two decades, only a few IBMAs have been conducted. Some researchers have leveraged IBMA methodologies for combining data from multiple sites from big data consortia, such as ABIDE[26] and the Brainnetome Project for Schizophrenia[27]. Other studies have followed a more traditional approach to meta-analysis by identifying relevant studies in the literature and then contacting the corresponding author of the selected publications, asking for the group-level unthresholded statistical maps[28–35]. Limiting IBMA to neuroimaging consortia restricts meta-analytic topics to a few domains and scientific questions. Alternatively, traditional meta-analysis is a more arduous task, as it requires contacting several dozen scientists to obtain their unthresholded and normalized statistical maps, which is time-consuming. Moreover, corresponding authors are not always responsive; the data could be lost, difficult to retrieve, or lack the quality required for a meta-analysis[28]. In comparison, more than 400 CBMA studies have been conducted since 2010, addressing consensus for many domains and scientific questions. In practice, researchers commonly report activation foci of significant findings in neuroimaging studies, making a large portion of the neuroimaging literature accessible for CBMA approaches[36–39]. Meanwhile, whole-brain statistical images are rarely shared, thus limiting IBMA to a small number of tasks and mental functions.

The NeuroVault web-based repository was introduced to address image data availability by providing an easy-to-use community platform to share statistical maps[4]. Today, NeuroVault's archives contain a significant volume of collections with more than two hundred thousand brain maps and corresponding metadata, some of which are linked to peer-reviewed articles. However, there is still limited coverage of the neuroimaging literature, as most research labs have not adopted NeuroVault as part of their workflows when submitting an article for publication[7]. Currently, downloading usable data from NeuroVault comes with multiple challenges. For example, a substantial portion of collections in NeuroVault are wrongly annotated or lack a link to a valid publication; some images are duplicates, and others correspond to non-statistical imaging modalities[6]. Overall, the potential number of spurious statistical maps complicates the use of NeuroVault data for IBMA. Although the NeuroVault API is integrated with some neuroimaging software (e.g., Nilearn and NiMARE)[7], users without coding experience often struggle to produce efficient analyses from NeuroVault data. Taken together, there is a need for standard guidelines for image selection and proper data cleaning, followed by open-source tools and reproducible methods to facilitate IBMA with NeuroVault.

The overall objective of the current study was to advance research tools for IBMA, including a standard image selection framework, image aggregation methods, databases, and guidelines. Initially, we examined the current status of the NeuroVault repository and evaluated its feasibility for IBMA. Then, we developed an image selection framework to identify images suitable for IBMA for a given domain or fMRI tasks. We implemented several combination methods with robust approaches to handle images with extreme values and outliers. The different combinations of image selection and combination methods were assessed against reference images from the task-fMRI group-average effect size maps from the Human Connectome Project (HCP) S1200 data release[40–44]. The comparisons between our IBMA results and the reference maps focused on evaluating image similarity and increased estimates in specific brain regions of interest. The entire process, from accessing data in NeuroVault to producing meta-analytic maps, is detailed in an open-access repository to facilitate reproducible and systematic meta-analysis. Collectively, we expect the results of the current work, including our specific guidelines, tools, and methods, to boost interest in the NeuroVault platform and promote IBMA research.

Results

Overview of the methodological approach

Figure 1 provides an overview of our methodological approach. First, we identified fMRI tasks linked to a specific domain (e.g., working memory) using the established connection between NeuroVault and the Cognitive Atlas knowledge base[4,45]. Then, we downloaded all images linked to selected tasks from the NeuroVault repository. Second, we performed a preliminary image selection leveraging the metadata associated with the images,

which were identified as potential candidates for IBMA. We also conducted a data-driven heuristic selection to remove possible outliers from the data. Subsequently, we manually selected relevant images by identifying the analysis contrast in their corresponding article and with the help of the image metadata in NeuroVault. Third, we conducted image-based meta-analyses using standardized effect size maps of the chosen images. In addition to using a baseline meta-analytic estimator (i.e., mean), also referred to in this article as combination or aggregation, we explored four robust combination methods: median, trimmed mean, winsorized mean, and weighted mean. Finally, the meta-analyses with different combinations of parameters (i.e., image selection method and estimator approach) were evaluated against reference images from the task-fMRI group-average effect size maps from the HCP S1200 data release[40–44]. Our proposed framework for IBMA using NeuroVault, along with a flexible Jupyter Notebook tutorial, is presented in the Methods section.

Exploring the current state of neurovault: data availability and quality assessment

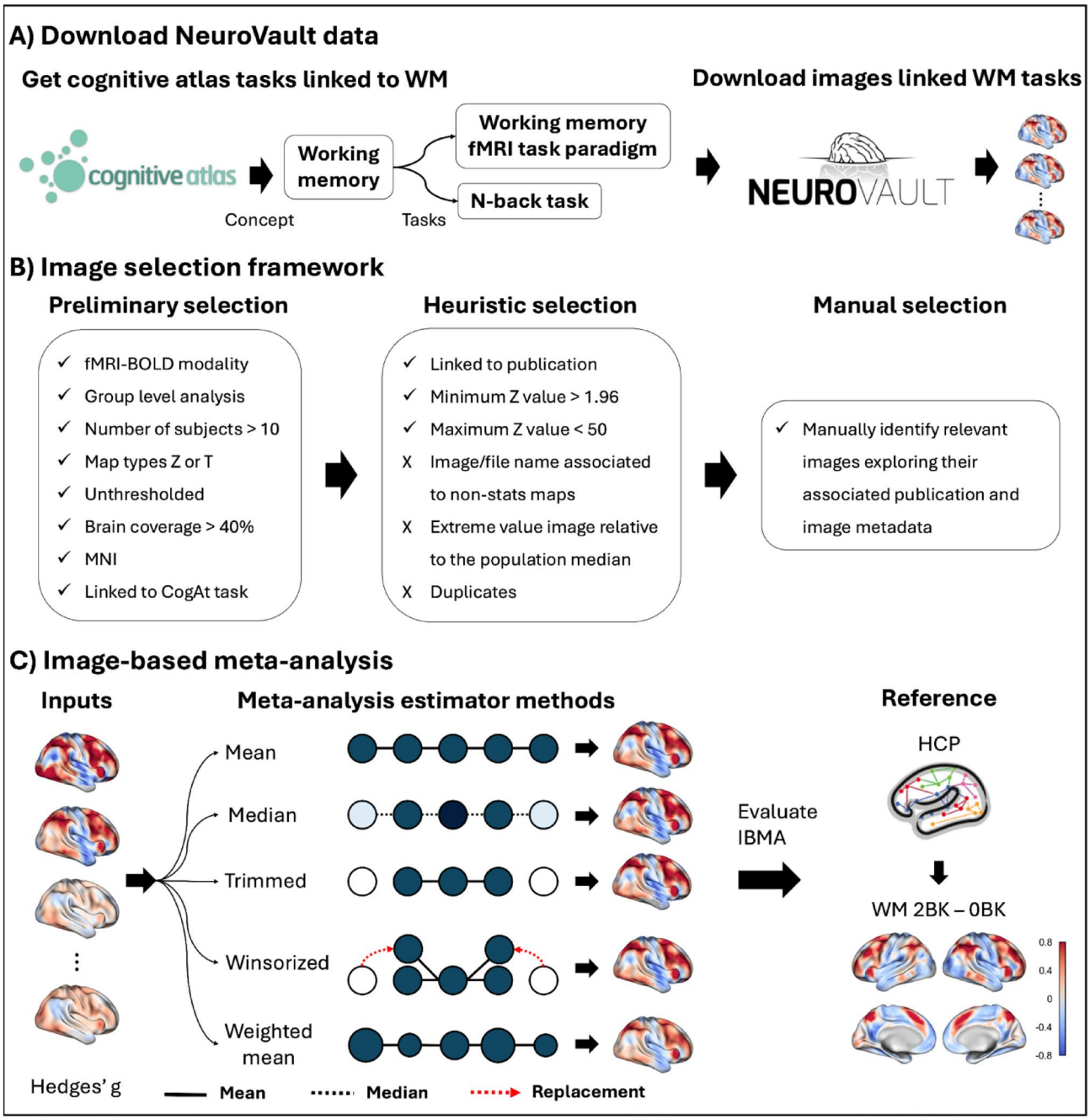
Before attempting to run IBMA, we conducted a thorough analysis of the state of NeuroVault as of February 2024 (Fig. 2). Figure 2A illustrates its evolution from 2013 to 2023. When NeuroVault was launched, it received 11 collections and 63 images in the first year. Since then, the amount of uploaded data has steadily increased, reaching over 238,000 images distributed across 5,756 collections a decade later. Of these collections, 26% are linked to published journal articles, and approximately 66% of images are associated with a publication. The growth of collections follows a similar trend, with a median of 618 new collections created each year. Notably, the most significant increases occurred in 2018 and 2023, with more than 700 collections added each year. The median number of images uploaded annually exceeded 11,000, with the highest uploads occurring in 2018, 2020, and 2021, totaling over 47,000, 54,000, and 58,000 images, respectively. Overall, these trends highlight the rapid growth and increasing adoption of NeuroVault as a comprehensive and widely used repository for sharing neuroimaging statistical images.

Next, we sought to identify which collections could be linked to publications with a valid DOI. Starting with NeuroVault's 5,756 collections, we found only 896 with a valid DOI link in the DOI field of the collection metadata. Of the remaining 4,985, we found 28 collections with a valid DOI link provided in the collection description field. We then searched PubMed for articles that matched the collection's title and found 354 additional collections that could be linked to published articles. Finally, we conducted an extensive search using Pubget[46], an open-source Python tool for collecting data for biomedical text mining. We performed a Pubget query and retrieved papers that mentioned NeuroVault in the title, abstract, keywords, or body ("*neurovault*[All Fields]") and found 194 additional collections linked to a valid publication. Altogether, we were able to identify a final sample of 1,472 NeuroVault collections that were linked to a published article with a valid DOI.

Importantly, not all collections and images in NeuroVault are "standard" collections uploaded by independent users and thus may not be suitable for meta-analysis. Specifically, 40% of the 9,502 collections were created by the Neuroscout web application[47], which utilizes NeuroVault to store results generated by its analysis pipelines for sharing and visualization purposes. Among the 5,756 remaining "standard" collections in NeuroVault, we examined how their images are organized based on modality, analysis level, and image type (Fig. 2B). As anticipated, the majority of images in NeuroVault are categorized as "fMRI-BOLD," with other modalities making up only 5% of the total. Interestingly, 25% of the images are not linked to existing image modalities. Notably, the distribution of images in NeuroVault is heavily skewed by a few large collections containing subject-level data. For example, collections 9494, 4337, and 8996 alone account for over 70,000 single-subject images, which skews the overall proportions. When examining collections linked to publications, we found that 797 collections contain only group-level data, while just 52 contain only single-subject data, and 18 contain both types. This distribution, with group-level collections outnumbering subject-level collections, clearly reflects NeuroVault's original purpose as a repository for sharing group-level statistical maps rather than individual subject data. Additionally, around 75% of the uploaded images are unthresholded, preserving the full richness of the data. Similarly, more than 75% of the images in NeuroVault are statistical maps (Z or T), with other types, such as variance and effect maps, representing a smaller proportion. Notably, 63% of the images are linked to valid Cognitive Atlas tasks.

NeuroVault images selection consideration for IBMA

The primary challenge with conducting IBMA using NeuroVault is identifying relevant images for a meta-analysis. To demonstrate this process, we conducted a preliminary selection of images that could potentially be suitable for meta-analysis (Fig. 3A). We focused on fMRI-BOLD images, as they are the most prevalent modality in NeuroVault, making up 70% of the total images. We specifically chose images from group-level analyses, which represent a smaller proportion compared to subject-level images, as illustrated in Fig. 2B. While using individual-level data and conducting a hierarchical IBMA mega-analysis represents the ideal scenario for data aggregation, such images in NeuroVault are concentrated in only a few large collections (e.g., collections 9494, 4337, and 8996 contain over 70,000 single-subject images combined), which restricts IBMA to just a few studies and neuroimaging domains. Additionally, information regarding the analysis contrast is not standardized in NeuroVault, making it impractical to curate such a large sample of images manually. In contrast, we focus on a more traditional IBMA using group-level statistical maps, which offers a more feasible approach for synthesizing findings across the existing NeuroVault data. The selection of group-level maps significantly reduced our sample to only 7.5% of all available images. However, the number of collections did not decrease at the same rate, reinforcing our observation that most single-subject images originate from just a few collections. Additionally, we retained only images from studies with a sample size greater than ten subjects. We selected images classified as T or Z statistics, bringing the total down to 11,422 images. Although best practices in meta-analysis suggest using meaningful units and incorporating uncertainty through standard errors, T/Z statistic maps are the most commonly shared images in NeuroVault. Moreover, we filtered for unthresholded images that covered at least



40% of the brain and were in MNI space. Finally, we narrowed our selection to 6,400 images associated with a Cognitive Atlas task, accounting for just 2.7% and 16.8% of the total image and collection sample, respectively. This selection process shows that only a small percentage of NeuroVault images are potentially relevant for image-based meta-analysis.

Following that, we aimed to explore how the Cognitive Atlas tasks were represented in the remaining images and collections (Fig. 3B). Social judgment and decision-making tasks had the most significant number of images, with 196 and 126 images distributed across 11 and 14 collections, respectively. The go/no-go task was included in the highest number of collections, with a total of 93 images across 17 collections. Other tasks that were notably well-represented in NeuroVault included motor, emotion processing, and n-back tasks, among others. Overall, a diverse range of tasks and domains is well represented in NeuroVault, providing sufficient data for conducting image-based meta-analyses.

For a preliminary evaluation of IBMA using NeuroVault data, we selected domains whose tasks are well-represented in NeuroVault. We focused on working memory, motor, and emotion processing. For working memory, we used the working memory fMRI task paradigm and the n-back task. For the motor domain, we selected images linked to the motor fMRI task paradigm, the motor sequencing task, and the finger tapping task. Finally, for emotion processing, the emotion processing fMRI task paradigm was considered. As the reference image for the three domains, we used effect size maps from the HCP. Additional analysis for other domains can be found in the supplementary materials (Fig. S1 and Fig. S2). In addition to the preliminary selection shown

◂**Fig. 1**. Systematic framework for conducting image-based meta-analysis using NeuroVault repository data. This three-stage workflow tackles the main challenge of systematically identifying and analyzing neuroimaging statistical maps from the NeuroVault database. (**A**) Data identification and acquisition: We utilized the established integration between NeuroVault and the Cognitive Atlas knowledge base to systematically find fMRI tasks within specific cognitive domains (e.g., working memory). Images and related metadata were then downloaded via the NeuroVault API for further analysis. (**B**) Multi-level image selection framework: Three complementary selection methods were applied to address data quality issues. Initial selection used strict inclusion criteria (fMRI-BOLD modality, group-level analysis, $N > 10$ subjects, T/Z statistical maps, unthresholded, > 40% brain coverage, MNI space) to pinpoint potentially suitable images for meta-analysis. Heuristic selection employed automated techniques to detect and remove spurious images, duplicates, and statistical outliers based on data-driven thresholds and filename patterns. Manual selection involved expert review of linked publications to verify contrast relevance and methodological correctness. (**C**) Robust meta-analytic estimation methods: Beyond standard mean aggregation, four robust estimators were used to manage extreme values and outliers often found in diverse datasets. The visualization shows how different estimators weigh individual studies: circle color intensity indicates relative influence in the final aggregation (darker = higher weight, white = excluded), circle size varies only for the weighted mean estimator (larger = higher precision), solid lines connect studies included in mean calculations, dotted lines represent median calculations, and red arrows indicate value replacement in winsorized approaches. All meta-analytic results were validated against high-quality reference maps from the HCP data to evaluate the effectiveness of various selection and estimation strategies.

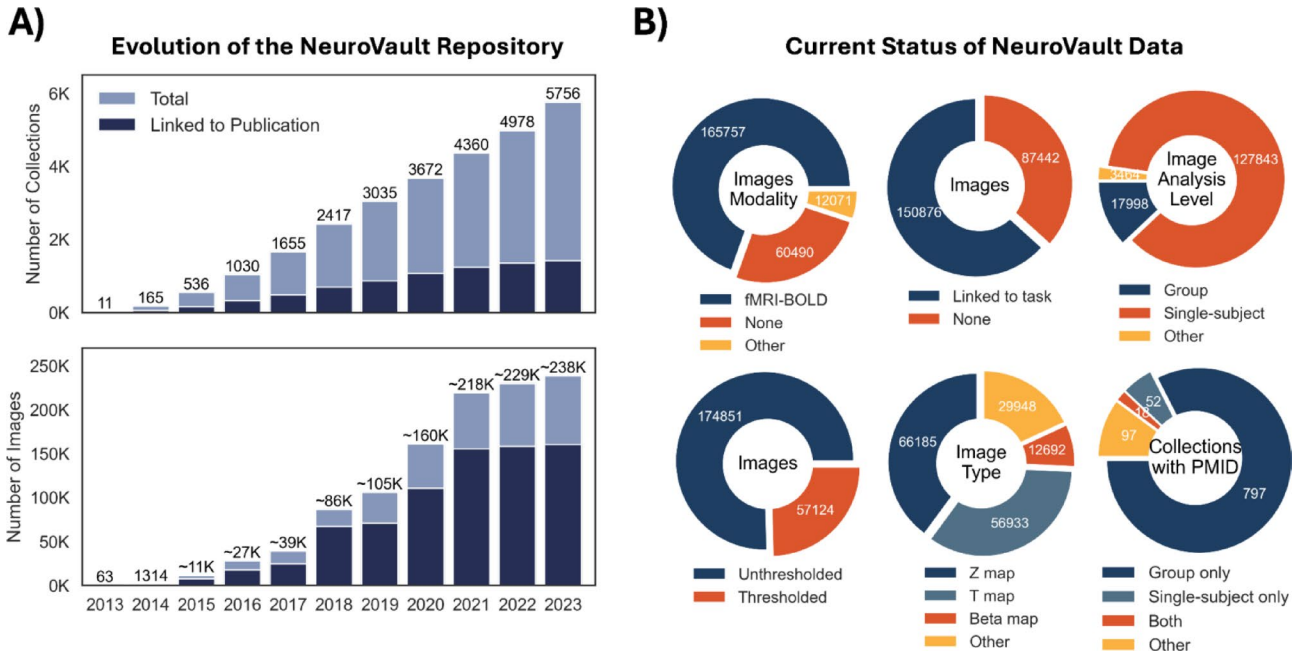


Fig. 2. Evolution and current state of NeuroVault data show rapid growth but limited meta-analysis-ready content. **(A)** NeuroVault's decade-long growth highlights increasing adoption of neuroimaging data sharing. The repository grew from 63 images at launch (2013) to over 238,000 images across 5,756 collections by 2023. Notably, only 26% of collections link to published articles (dark blue bars), revealing a challenge for systematic meta-analyses. The steepest growth occurred between 2018 and 2021, with annual uploads surpassing 47,000 images. **(B)** Distribution of NeuroVault content points out key limitations for IBMA. **Top row**: While fMRI-BOLD is the most represented modality (70% of images), 25% lack modality specification. Only 63% of images are linked to standardized Cognitive Atlas tasks, which are essential for domain-specific meta-analyses. Subject-level data dominate the analysis-level distribution (top right), but the overall proportions are skewed from just a few large collections. Bottom row: Encouragingly, 75% of images are unthresholded, preserving the statistical information necessary for IBMA. Statistical map types (Z/T) are predominant. The collection-level view (bottom right) shows that among publications with PMIDs, group-only collections ($n = 797$) greatly outnumber single-subject collections ($n = 52$), aligning with NeuroVault's intended use for sharing group-level results and supporting our focus on traditional IBMA approaches.

in Fig. 3A, we excluded images from the HCP NeuroVault collection. Consequently, our criteria resulted in 98 images distributed across 19 collections for working memory, 85 images in eight collections for motor, and 82 images in 14 collections for emotion processing (Table S1). This set of images is referred to as "All Images" throughout this paper.

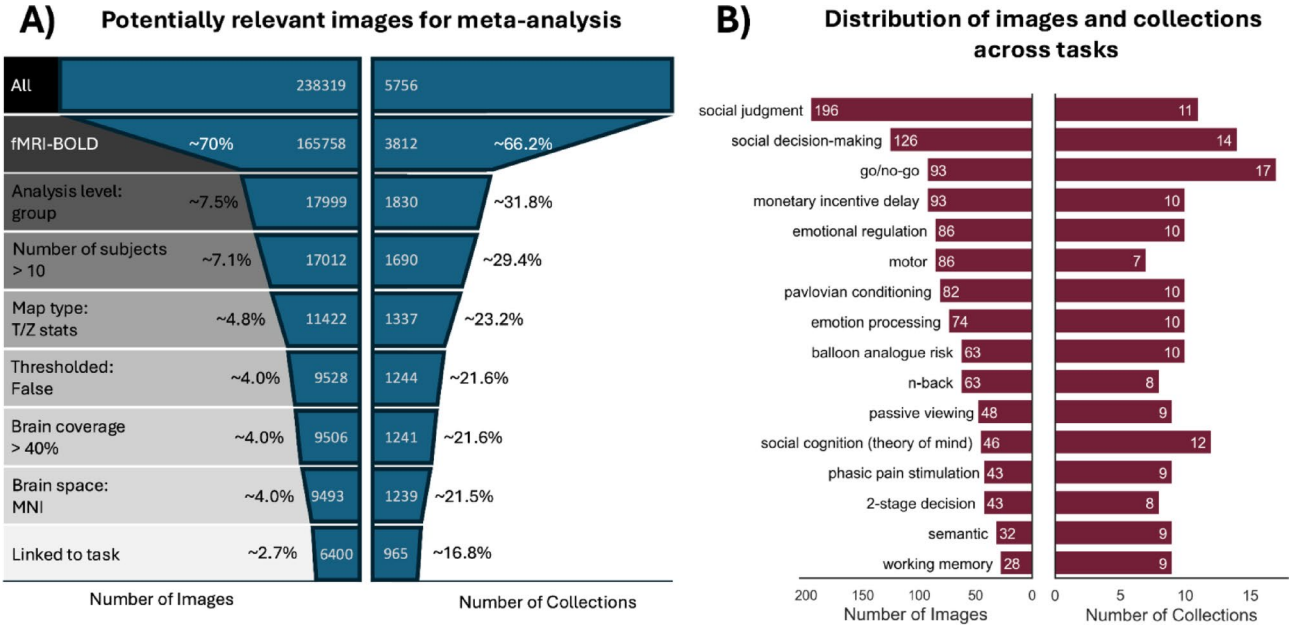


Fig. 3. Systematic filtering shows that only 2.7% of NeuroVault images meet preliminary IBMA criteria, with enough coverage across cognitive domains. **(A)** Applying standard neuroimaging meta-analysis criteria significantly reduces the usable NeuroVault dataset. Starting from 238,319 total images in NeuroVault, each filter removes unsuitable data: limiting to fMRI-BOLD (70%), group-level analyses (7.5%), sample sizes (> 10 subjects), statistical maps (T/Z), unthresholded data, sufficient brain coverage (> 40%), MNI space normalization, and Cognitive Atlas task annotation. This strict but necessary selection process results in only 6,400 potentially usable images (2.7%). The most restrictive criterion is the group-level requirement, indicating that most NeuroVault data comes from individual subject uploads rather than study-level summaries. **(B)** The distribution of these 6,400 filtered images across cognitive domains shows enough representation for targeted meta-analyses. Social judgment and social decision-making tasks have the most images (196 and 126, respectively), while go/no-go tasks cover the most collections (17). Importantly, our three focus domains (i.e., working memory (n-back), motor tasks, and emotion processing) all have sufficient data for solid IBMA. The dual-axis chart reveals that image count and collection diversity don't always match; some tasks with many images are concentrated in a few collections, which could limit generalizability. This distribution confirms NeuroVault's potential for domain-specific IBMA despite the strict filtering.

After selecting these well-represented and exemplar domains, we applied data-driven heuristic selection methods that primarily focused on identifying images characterized by extreme values, duplicates, and inverted contrasts. For example, we only selected images with a minimum Z value greater than 1.96 and a maximum Z value less than 50. We eliminated images associated with non-statistical maps by detecting patterns in the image and file names, such as "ICA," "PCA," and "PPI," among others. Consequently, we removed extreme images relative to a robust average of the whole population of images using regression slopes. This heuristic selection framework reduced the number of collections and images in domains such as working memory and emotion processing by half. However, for the motor domain, only one collection and a few images were removed (Table S1).

Finally, we manually selected the most relevant images of the publications linked in the NeuroVault collections. Our focus was primarily on the task description outlined in the paper's method section and the specific contrast of interest. To aid in our selection, we examined the image, file name, and contrast definition fields found in the image metadata within NeuroVault. Ultimately, the final sample comprised ten images from six collections for working memory, 30 images from seven collections for motor, and eight images from just two collections for emotion processing (Table S1).

Evaluating manual IBMA with NeuroVault

To evaluate IBMA using NeuroVault, we conducted a meta-analysis on the manually selected images corresponding to the three domains described above. The input images, initially downloaded as T/Z statistics, were converted into Cohen's d maps using the sample sizes available in their NeuroVault metadata. We employed a baseline IBMA estimator (i.e., the mean), which performed a voxel-wise average of the input maps. For reference maps, we utilized the group-average effect size maps from the HCP S1200 data release[40–44]. Specifically, for working memory, we used the contrast "2-Back vs. 0-Back"; for motor, we aimed to reproduce the contrast representing the average of all motor movement blocks against the baseline "Motor vs. Baseline," and for emotion processing, we employed the contrast "Face vs. Shape." Additional details regarding these tasks and their available contrasts are described by Barch et al.[40]. We compared the results of our manual IBMA using the baseline estimator and reference maps from the HCP data (Fig. 4). HCP group-average maps were chosen as reference standards for

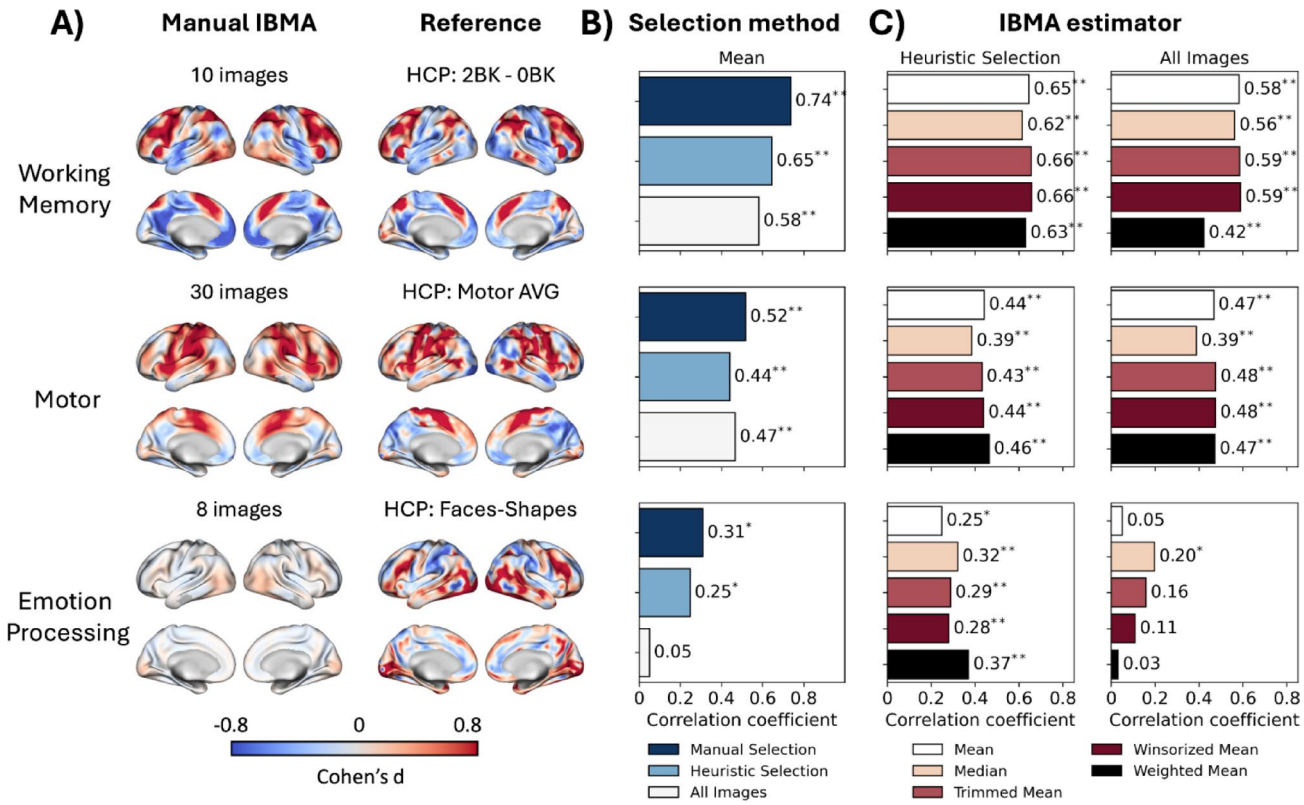


Fig. 4. Manual curation yields the strongest meta-analytic results, but automated methods with robust estimators provide viable alternatives for heterogeneous domains. **(A)** Visual comparison shows that manual IBMA successfully recovers activation patterns from HCP reference data. Working memory IBMA captures bilateral frontoparietal networks typical of n-back tasks. Motor IBMA accurately identifies primary motor cortex activation. Emotion processing shows weaker but anatomically plausible visual cortex activation, reflecting greater heterogeneity in face-processing paradigms across studies. Color scales represent effect sizes, with warmer colors indicating stronger activation. **(B)** Quantitative validation reveals how image selection strategies impact IBMA. Manual selection consistently outperforms automated approaches, achieving the highest correlations with HCP references (working memory: $r = 0.74^{**}$, motor: $r = 0.52^{**}$, emotion: $r = 0.31^{*}$). Heuristic selection, which automatically removes outliers and spurious images, improves results over including all images, especially for emotion processing. The small difference between methods for motor tasks ($\Delta r = 0.05$) suggests that homogeneous paradigms require less curation. At the same time, emotion processing shows significant improvement with manual selection ($\Delta r = 0.26$). **(C)** Robust statistical estimators show domain-specific benefits when manual curation is not feasible. For homogeneous domains (i.e., working memory or motor), all estimators perform similarly, indicating traditional mean-based approaches are sufficient. However, for heterogeneous emotion processing data, robust methods significantly enhance results: median estimation increases correlation from $r = 0.05$ to $r = 0.20^{*}$ when all images are included, approaching manually curated results. The weighted mean shows high variability: best with heuristic selection ($r = 0.37$) but worst with unfiltered data. Statistical significance was tested with spin permutation tests that account for spatial autocorrelation ($^{*}p < 0.05$; $^{**}p < 0.001$).

validation, given their coverage of distinctive domains, scale, and quality of the data. We tested whether our IBMA can detect robust, reproducible task effects despite methodological heterogeneity, rather than expecting identical results. High correlations indicate successful identification of core task-related patterns; lower correlations may reflect either limitations in the IBMA approach or heterogeneity in the literature. The degree of correspondence informs us about both the quality of available data and the effectiveness of different selection and aggregation strategies.

Manual IBMA effectively reproduced the reference maps for the three domains, demonstrating significant qualitative and quantitative convergence (Fig. 4A). Notably, the working memory and motor domains exhibited estimated effect sizes comparable to the reference maps. In contrast, while there were similarities observed in visual brain regions for emotion processing, the average effect size estimate was weaker. Overall, these findings indicate that the NeuroVault data is suitable for IBMA.

Assessing the impact of image selection and estimator methods on IBMA

Despite the availability of relevant data in NeuroVault, conducting a manual image-based meta-analysis still requires significant effort, as described above. Moreover, the current process for selecting relevant images from

NeuroVault is not standardized, which may lead to biases introduced by researchers. With this in mind, we next considered the feasibility of applying a heuristic to selecting relevant images or including all images in a meta-analysis without the extensive manual work. Manual selection continues to be the preferred method for meta-analyses where accuracy is crucial. Our review of automated methods aims to achieve two goals: (1) to check if automated techniques can produce acceptable initial results for exploratory studies or screening many potential domains; and (2) to support NeuroSynth-style automated meta-analyses that can monitor how findings develop as NeuroVault expands. In this section, we investigate if these alternatives to the manual approach yield acceptable results.

In Fig. 4B, we compared correlation coefficients from each meta-analysis based on three selection methods: (i) one that included all images from the preliminary selection, (ii) another that used a data-driven heuristic selection, and (iii) a third that relied on manual selection. As anticipated, the manual meta-analyses yielded the highest correlation with the reference maps across the three domains. Specifically, the manual IBMA for working memory showed a strong correlation ($r = 0.74$, $p < 0.001$), followed by the motor domain ($r = 0.52$, $p < 0.001$), and a weaker correlation for emotion processing (0.31, $p < 0.001$). Surprisingly, including all images in the meta-analysis produced satisfactory results, with correlations of $r = 0.58$ ($p < 0.001$) for working memory and $r = 0.47$ ($p < 0.001$) for the motor domain. Including all images resulted in almost no correlation for emotion processing. The difference between the manual selection and the all-images approach for the motor domain was relatively small, with a 0.05 increase observed with manual selection. However, for the other two domains, the difference was more significant: there was a 0.16 gain for working memory and a 0.26 gain for emotion processing using the manual approach. Finally, the heuristic selection method showed promising results for working memory and emotion processing, suggesting that this approach effectively minimized extreme and unwanted values from the all-images sample. In fact, the correlation coefficients for these two domains were more aligned with those found through manual selection. In the case of the motor domain, the heuristic selection did not eliminate many maps and maintained the same number of collections (Table S1); as a result, the correlation achieved through the heuristic meta-analysis was slightly lower than that obtained with the all-images approach.

Meta-analysis methods can be notably influenced by extreme observations. In most cases, except for random-effects meta-analysis, an outlier is viewed as strong evidence of an effect and can skew results toward significance. Conversely, in a random-effects meta-analysis, an outlier can increase heterogeneity and ultimately reduce the significance of the findings. Here, we propose four robust alternative methods to the baseline mean (Fig. 4C). Using the median, we achieve maximum robustness even when up to half the data is corrupted, although it is somewhat less efficient than the mean in terms of variance. The trimmed mean removes the most extreme 10% of observations from each tail, providing a balance between robustness and efficiency by excluding outliers. The winsorized mean replaces extreme values with their nearest neighbors instead of removing them entirely, maintaining the original sample size while lowering the impact of outliers. Lastly, the weighted mean uses inverse-variance weighting, assigning greater influence to studies with smaller standard errors.

For domains characterized by tasks with more homogeneous experimental paradigms, such as working memory and motor, the different estimators correlated similarly with the reference maps. Estimators that specifically target extreme values in the data, like the trimmed mean and winsorized mean, performed slightly better than the baseline estimator. However, the increase in correlation of only 0.01 does not justify the need to modify or reduce the input sample using those robust estimators. For the motor tasks, the median estimator performed worse than the other estimators, while the weighted mean showed relatively strong performance. In the case of working memory, the weighted mean yielded the best performance when including all images; this may be attributed to the presence of outlier studies with small standard errors. For a domain defined by tasks with high heterogeneity of experimental paradigms, such as emotion processing, robust estimators significantly improved upon the baseline. Interestingly, when all images were included, we initially found a correlation of $r = 0.05$ ($p = 0.678$). However, using the median as a combination method increases this correlation to $r = 0.2$ ($p = 0.041$), which was nearly on par with the results obtained through heuristic selection using the baseline estimator. All robust estimators enhanced the correlation when applied to the heuristic selection sample, with the weighted mean achieving a correlation of $r = 0.37$ ($p < 0.001$), surpassing the correlation found in the baseline manual meta-analysis. Yet, the weighted mean performed the poorest when all images were considered in the meta-analysis. Overall, the trimmed mean provided the most consistent improvement across different image selection methods and domains. In manual IBMA, robust estimators performed similarly to the baseline estimator, with the mean estimator occasionally showing slightly better results. The weighted mean tended to produce the most extreme values, significantly increasing the correlation in some cases while decreasing it in others compared to the reference. Additional results of all parameter combinations tested are available in the supplementary materials (Fig. S3, Fig. S4, Fig. S5, Fig. S6, Fig. S7, Fig. S8, Fig. S9, Fig. S10, Fig. S11, Fig. S12, Fig. S13, and Fig. S14).

Following this, we aimed to investigate how the estimated effect varied with different image selection methods. Figure 5 presents the distribution of estimates for the three domains, categorized by image selection methods. Not surprisingly, the results of the manual meta-analysis demonstrated broader distributions centered around zero, which indicates a relative increase in both positive and negative effect size estimates. Notably, for the motor domain, the distribution resulting from manual selection appeared more positively skewed than the other selection methods. In contrast, when we included all images in the meta-analysis, we observed the narrowest distributions among all selection methods. The distributions generated by applying the heuristic selection were slightly wider than those produced by including all images. These findings are qualitatively supported by the activation levels in the surface map plots, which clearly show that manual selection results in stronger effect estimates than the other methods. Meanwhile, the heuristic selection method still enhances the results compared to incorporating all images in the meta-analysis.

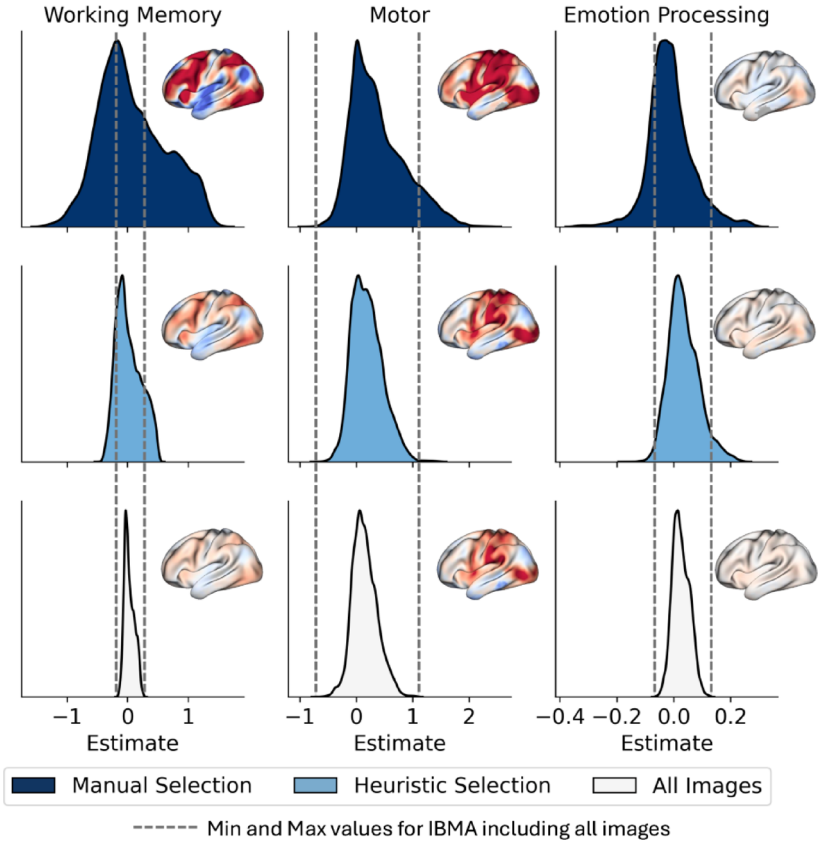


Fig. 5. Manual selection produces stronger effect sizes, revealing signal reduction when including potentially heterogeneous images. The distribution of vertex-wise effect size estimates shows how image selection methods influence meta-analytic sensitivity and specificity. Manual selection (top row, navy) results in the broadest distributions with prominent positive tails. This is especially clear for motor tasks, where manual curation creates a noticeable rightward skew, reflecting robust activation in the motor cortex. Conversely, including all images (bottom row, gray) leads to narrow, zero-centered distributions, with signs of signal reduction caused by heterogeneous contrasts, spurious images, and inverted maps that cancel out true effects. Heuristic selection (middle row, light blue) yields intermediate results, partially recovering signal strength by automatically filtering outliers and duplicates. The impact varies across domains: working memory and motor tasks show noticeable improvements with curation (notice the extended positive tails). At the same time, emotion processing remains near zero across all methods, highlighting the heterogeneity of the contrasts in NeuroVault. Vertical gray lines indicate the range from the "All Images" selection, emphasizing how manual selection broadens the dynamic range of detected effects. Surface visualizations (scaled − 0.8 to 0.8) confirm that wider distributions correspond to stronger, more focused activation patterns.

Subsequently, we quantitatively evaluated the increased estimated values in specific areas of interest across the various domains (Fig. 6). To achieve this, we focused our analysis on particular regions of interest, defined by selecting the top 10% of vertices from the reference maps (Fig. 6A). The results of this evaluation for all image selection methods are displayed in Fig. 6B. As anticipated, the manual meta-analysis produced the highest estimates, approaching one, for working memory and motor. Heuristic selection slightly improved the estimates compared to using all images, yielding mean estimates of 0.31 for working memory and 0.49 for motor. In contrast, including all images resulted in the lowest mean estimates across all selection methods. Interestingly, the estimates for emotion processing showed no significant differences among the selection methods, with all mean values close to zero. Regarding the meta-analytic estimators, the baseline estimator outperformed the four robust estimators for both heuristic and all image selection methods in the cases of working memory and motor. The median and weighted mean estimators exhibited the poorest performance among the five estimators. Overall, the five methods produced similar average estimates for emotion processing, with values just above zero.

Discussion

The NeuroVault repository currently houses a vast collection of brain images. However, the absence of standardized methods, accessible tools, and clear guidelines has hindered the reusability of this rich data source for secondary data analyses, such as IBMA. In this study, we examined the current state of the NeuroVault data, developed an automatic heuristic for selecting images, implemented five estimation methods, and provided guidelines for conducting IBMA. First, we highlighted the rapid growth of NeuroVault as a comprehensive repository for sharing

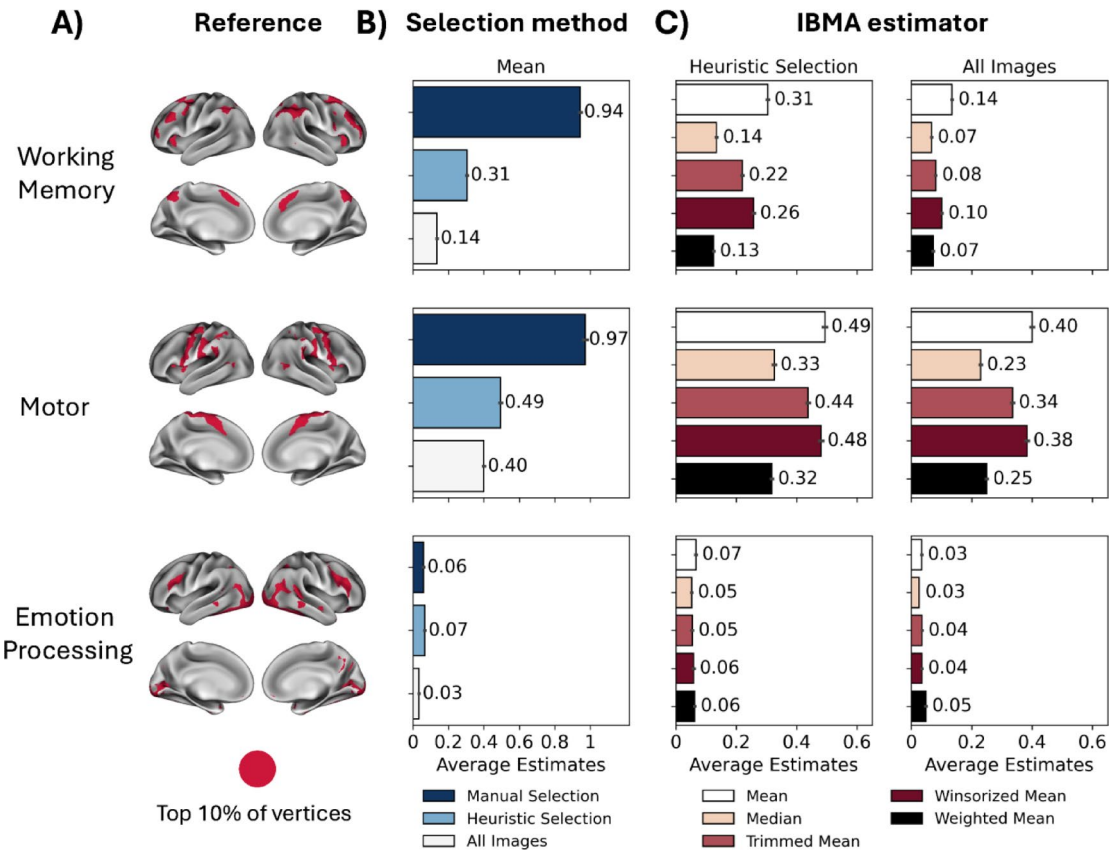


Fig. 6. Effect sizes in task-relevant regions confirm that manual selection is superior, with domain-specific patterns revealing essential differences in task uniformity. **(A)** Reference regions of interest (ROIs), defined by the top 10% of effect sizes from HCP data, capture typical activation patterns for each cognitive domain. These ROIs serve as targeted benchmarks for assessing meta-analytic recovery of task effects. **(B)** Quantification within reference ROIs shows substantial differences in meta-analytic sensitivity depending on the selection method. Manual selection nearly maximizes recovery for homogeneous domains (working memory: $d = 0.94$, motor: $d = 0.\ 0.97$). Heuristic selection offers a moderate improvement over unfiltered data (working memory: $d = 0.31$ vs. 0.14, motor: $d = 0.49$ vs. 0.40), suggesting that automated outlier removal partially addresses data quality issues. Notably, emotion processing shows negligible effects ($d \approx 0$) regardless of the selection method, implying that data and contrast heterogeneity fundamentally limit effect size recovery even in task-relevant regions. **(C)** Robust estimators provide little benefit for automatically curated data but cannot fix heterogeneous datasets. For heuristic selection, traditional mean estimation remains best, while robust methods (especially median and weighted mean) can reduce effect sizes by 20–50%. When all images are included, no estimator significantly improves emotion processing results.

neuroimaging statistical images. We noted a diverse representation of tasks and domains supporting secondary data analyses. Second, we demonstrated the feasibility of conducting IBMA with NeuroVault by reproducing reference maps from the HCP data for three distinct domains: working memory, motor, and emotion processing. We emphasized the importance of proper image selection for IBMA, particularly for heterogeneous domains such as emotion processing. Finally, we implemented robust IBMA estimators to manage extreme or spurious images within the input sample, underscoring their relevance for more complex domains.

Present and future of NeuroVault data for IBMA

NeuroVault has been widely adopted by researchers in the neuroimaging community. In just ten years, it has grown from 61 images at its launch in 2013 to over a quarter of a million images today. Our systematic analysis revealed significant challenges within the NeuroVault repository. Despite hosting over 238,000 images across 5,756 collections, only a small fraction were potentially relevant for IBMA. Specifically, only 2.7% of the images met our preliminary inclusion criteria, where requiring group-level images for meta-analysis represented the strictest criteria. This low percentage is primarily explained by the fact that subject-level data from just a few extensive collections dominate the total image count. However, when examining collections rather than raw image counts, we found a reasonable representation of group-level analyses across NeuroVault, consistent with the repository's original purpose of sharing group-level statistical maps. The plateau shown in Fig. 2B in publication-linked collections in recent years may suggest a potential decline in the availability of imaging data for meta-analysis in the future. However, we believe this pattern is likely due to publication time lags, and we

expect that more existing collections will eventually be linked to publications. While some collection owners may not add the publication DOI to their collections post-publication, we can effectively search PubMed for any publications missing from the NeuroVault collections. In summary, our results address previous concerns about the feasibility of IBMA concerning data availability. We have shown that NeuroVault contains enough data to conduct IBMA despite strict filtering criteria. The images suitable for meta-analysis cover a diverse range of domains, including social judgment, decision-making, motor, emotion processing, and working memory.

Our comprehensive analysis of the NeuroVault repository revealed that many collections and images face data quality issues, such as incorrect annotations, missing publications, and non-statistical maps. This underscores the urgent need for standardized data curation methods to facilitate secondary analysis using NeuroVault data. While we initially conducted a manual meta-analysis to address data quality concerns, this approach is labor-intensive. Additionally, the process of manually selecting relevant images lacks standardization, which can lead to inconsistencies and reproducibility issues across studies. To tackle these challenges, we developed and validated an automated framework for removing potentially spurious images and other outliers. This framework allows researchers to efficiently identify and filter suitable statistical maps for IBMA, thereby promoting good practices and guiding future meta-analyses. Our method has enhanced reliability and reproducibility, and case studies demonstrated how effective data filtering and model selection can significantly influence meta-analytic results. Our findings indicate that, while manual selection remains optimal for detecting stronger effect sizes and achieving broader activation distributions, combining heuristic selection with appropriate estimators could provide a practical balance between accuracy and efficiency in image-based meta-analysis. The effectiveness of robust estimators varied by domain, with the trimmed mean displaying consistent performance across different selection methods. In general, we are most confident in our results when all these robust methods give similar answers and agree with the conventional mean (e.g., the results for the heuristic selection for working memory and motor). When these robust methods dramatically differ from the mean (e.g., the results of including all images for emotion processing), this indicates that extreme observations play an appreciable role. Unfortunately, though, the reduced efficiency implies that with a very small sample size, these robust methods could produce rather extreme results.

In summary, our study emphasizes both the potential and the limitations of IBMA using NeuroVault. Although the repository holds over 200,000 images, a significant gap exists between ideal meta-analysis practices and the available data. IBMA mega-analysis using individual participant data remains the gold standard, offering the best statistical power. However, subject-level data are mainly available in a few large collections, making large-scale mega-analyses across different domains impractical. This pattern reflects broader data sharing practices: while some large projects (e.g., HCP collection) provide extensive individual datasets, most researchers share only group-level statistical maps, if any. NeuroVault was originally designed for sharing group-level data due to privacy concerns, storage limitations, and its initial purpose of conducting meta-analyses with group data[4]. This explains why, although about 10% of images are group-level analyses, these are dispersed across many collections. Additionally, the lack of standardized contrast annotations for individual data makes manual curation difficult, limiting researchers to traditional IBMA methods. Despite these limitations, our results demonstrate significant improvements over coordinate-based approaches. Moving forward, the community should encourage individual data sharing through existing platforms, even without perfect harmonization techniques, as this would increase reuse potential. Our framework provides a practical solution for the current landscape and highlights areas where infrastructure improvements are needed. Future efforts should focus on developing automated annotation methods, improving the quality and coverage for both subject-level and group-level data.

Practical considerations for researchers

While our framework provides systematic methods for conducting IBMA with NeuroVault data, researchers must carefully address several methodological issues to ensure valid and reproducible results. First and foremost, the assumption of independence between studies is essential for accurate meta-analytic inference. We found that multiple images from the same collection or study can easily violate this assumption, potentially inflating false positive rates. Researchers should thoroughly document when multiple images from a single study are included and consider using hierarchical models that account for such dependencies[48]. Additionally, contrast heterogeneity presents a major challenge; for example, in heterogeneous domains like emotion processing, the specific contrasts can vary considerably (e.g., faces vs. shapes, emotional vs. neutral faces, fear vs. happy expressions). Manual inspection and careful selection of truly comparable contrasts are vital, as our results show that heterogeneous contrasts significantly decrease meta-analytic sensitivity.

Our analysis showed that metadata accuracy in NeuroVault cannot be assumed. We found many instances of mislabeled statistical map types, incorrect modality specifications, and missing or inaccurate sample sizes. Cross-referencing with original publications was essential but time-consuming, highlighting the need for better data curation standards. The issue of sample size reporting needs special attention, as this information directly impacts meta-analytic weighting schemes. When sample sizes were missing or questionable, we had to make assumptions that could have influenced our results, such as removing those studies from the meta-analysis. Visual inspection of statistical maps, even after automated quality control, revealed additional spurious images that could have contaminated the meta-analysis. These quality issues emphasize that automated methods, while useful for efficiency, cannot fully replace careful manual review. Our results confirm that manual selection yields better results, especially for domains with heterogeneous experimental paradigms. Nonetheless, we show that heuristic selection can be a reasonable compromise in several situations: (1) during preliminary or exploratory analyses, (2) when screening multiple domains to identify those needing detailed manual review, (3) working with well-defined, homogeneous experimental paradigms (like a motor task), or (4) when resources for manual review are limited (for example, if there is no link to the publication in the selected image).

To facilitate practical implementation, we suggest starting with the Cognitive Atlas to systematically identify relevant tasks and domains, downloading related images and collections from NeuroVault, and then thoroughly reviewing linked publications to understand methodological details and verify contrast comparability. Although this workflow is currently manual, it is being integrated into Neurosynth Compose to make the process more efficient[1]. Nevertheless, even with better tools, the core issues of data quality, independence, and heterogeneity will demand ongoing vigilance from researchers performing IBMA with shared neuroimaging data.

Limitations and challenges

Several limitations in the current work must be highlighted to move the field forward. The meta-analytic estimator method used in this work solely focuses on handling extreme and unwanted observations. We considered all images to be independent of each other. Nevertheless, in most cases, NeuroVault images belong to the same paper or collection. Thus, they cannot be considered independent since they may undergo the same analytic pipeline or could correspond to the same population. Critically, our assumption of independence can lead to inflated false-positive results. A recent work on the same data meta-analyses has studied this issue, proposing multiple models to account for dependencies[48]. These models were evaluated for Stouffer's and the generalized least-squares problem. Future work is required to incorporate and test these models in our robust estimator methods.

The automated analysis using the heuristic selection approach might combine images from different contrasts or experimental conditions. This variability can hide or confuse meta-analytic results. Additionally, the inclusion criteria in the preliminary and heuristic selection depend on metadata provided by NeuroVault contributors. Misannotations or incomplete metadata (such as incorrect map type, modality, or sample size) can lead to improperly including or excluding images. Even after automated outlier detection and quality control, some inaccurate, mislabeled, or low-quality images may still be present, potentially biasing the results. Moreover, not all images have accurate or available sample size data, which can affect their weighting in meta-analytic estimators that rely on sample size. We recommend researchers see automated selection as a supplement to, rather than a replacement for, careful manual curation. The heuristic method can help identify candidate images for review, alert to potential issues, and provide initial results to inform more detailed analysis.

We only included analysis for three domains as a proof of concept. Although additional analyses were conducted and presented as supplementary information, such results could not be thoroughly evaluated, given the lack of reference maps from large-scale fMRI studies. Moreover, other domains could not be considered because of data availability. However, we are confident that, as NeuroVault repositories continue to grow, our observations will be evaluated at a larger scale. The results presented in this paper, along with the methodology and tools, will be critical in extracting meaningful scientific insights from this increasingly large and complex database.

While we used HCP group-average maps as our main reference, we recognize that comparing our results to other published IBMA would have offered additional validation. However, to our knowledge, no previous IBMA has been conducted for the broad cognitive areas examined here (working memory, motor, and emotion processing as general categories). Previous IBMAs have generally focused on more specific contrasts or clinical populations (e.g., Schulze et al., 2016[34] on emotional face processing in bipolar disorder; Lukow et al., 2021[32] on reward processing). The absence of comparable broad-domain IBMA highlights the novelty and significance of our contribution in demonstrating the feasibility of domain-level meta-analyses using NeuroVault.

Conclusions

This study advances neuroimaging research by providing a comprehensive, reproducible framework for conducting IBMA using NeuroVault data. Our findings highlight both the challenges and potential of the NeuroVault repository. The methodology presented here offers researchers a robust set of tools and methods for assessing data quality, implementing flexible image selection strategies, and conducting reliable meta-analyses across diverse domains. Future work should focus on further refining automated selection techniques, expanding the range of domains analyzed, and developing more robust estimation methods. As the NeuroVault repository continues to grow, such standardized approaches will be critical in extracting meaningful scientific insights from the increasingly large and complex neuroscience literature.

Methods

Databases

NeuroVault

NeuroVault (https://neurovault.org) is a web-based repository of fMRI statistical maps from neuroimaging studies[4]. The brain maps are grouped in collections that are created and updated voluntarily. This repository can be explored and downloaded with the help of an API, which is supported by some Python neuroimaging tools (e.g., Nilearn and NiMARE). As of January 2024, NeuroVault contained 238,319 maps distributed in 5,756 collections, of which approximately 1,473 were associated with a published paper. The collections were linked to papers using the DOI field and collection description from their metadata. We also searched PubMed for articles that matched the collection's title. Additionally, we conducted an extensive search using Pubget, an open-source Python tool for collecting data for biomedical text mining (https://neuroquery.github.io/pubget/pubget.html). We performed a query and retrieved papers that mentioned NeuroVault in the title, abstract, keywords, and body of the articles ("*neurovault*[All Fields]").

To explore the NeuroVault database, we created an SQL query and exported the database contents to human-readable tables while filtering sensitive user information. This provided sufficient metadata from all collections and images to investigate the entire database without downloading the files. The images identified as usable for

IBMA (see the following section on the image selection framework) were downloaded along with their metadata and converted to a NiMARE Dataset object to leverage existing IBMA methods implemented in NiMARE.

Cognitive Atlas

Cognitive Atlas[45] (https://www.cognitiveatlas.org/) is an online repository of cumulative knowledge from experienced researchers from the psychology, cognitive science, and neuroscience fields. The repository currently offers two knowledge bases: 907 cognitive concepts and 841 tasks with definitions and properties. Cognitive concepts contain relationships with other concepts and tasks to establish a map between mental processes and brain function. It provides an API to download the database, which is also integrated into NiMARE.

NeuroVault image selection framework

Preliminary selection

Using the available metadata from the retrieved tables, we set different inclusion criteria for images to be considered for a meta-analysis. We focused on fMRI-BOLD images, as they are the most prevalent modality in NeuroVault. Note that the methods presented in this paper should work with other image modalities (e.g., PET, diffusion MRI, structural MRI). Still, only fMRI-BOLD had enough data in NeuroVault for meta-analyses. Then, we specifically chose images from group-level analyses. Additionally, we retained only images from studies with a sample size greater than ten subjects. Next, we selected images classified as T or Z statistics. Although best practices in meta-analysis suggest using meaningful units and incorporating uncertainty through standard errors, T/Z statistic maps are the most commonly shared images in NeuroVault[49]. We discuss this further in the following sections. Upon review, it is essential to note that many images in NeuroVault are labeled as "Other" for the image type. Nonetheless, most of those images actually correspond to known image types (e.g., T/Z statistic). As a result, we relabeled those images to their original type if keywords such as "zstat," "tstat," "Z_," or "T_" were present in the image name, file name, or image description. Following that, we retained unthresholded images that cover 40% of the brain and are in MNI space. Ultimately, we narrowed our selection to images associated with a Cognitive Atlas task.

Heuristic selection

Even after applying the previous strict preliminary inclusion criteria, we still found plenty of wrongly annotated images, especially representing other image modalities and others with extreme values. Therefore, we developed an automatic heuristic selection to remove those spurious images from the meta-analysis. The heuristic selection consisted of two steps. First, we removed all images from collections that lacked a link to a publication. Also, images with a minimum Z value smaller than 1.96 (i.e., Z score for a 0.05 p-value) were removed as they potentially consisted of mislabeled correlation maps, inverted p-value maps, or did not contain statistically significant voxels. We also excluded images with a maximum Z score larger than 50. Although the number 50 is arbitrary, we wanted to detect images with an unusually large signal. For example, mislabeled BOLD or COPE (contrast of parameter estimates) images or others resulting from studies with a large sample size. Additionally, using the image metadata, we analyzed the image and file name. We removed those containing keywords such as "ICA," "PCA," "PPI," "seed," "functional connectivity," "cope," "tfce," and "correlation," which represent modalities not of interest for the meta-analysis of the current work.

Second, we aimed to detect and remove extreme images in relation to a robust average of the entire image population. Note that the population of images to calculate the average was considered on a domain basis, and not the entire sample of images from NeuroVault. After creating the robust average images (i.e., the median), we made a rough segmentation of 'signal' and 'noise' voxels. For signal, we defined the mask as the bottom and top 10% of voxels (by rank order); for 'noise,' conversely, we selected the 20% of voxels with the smallest magnitude (i.e., closest to 0). Then, we performed the correlation exercise only on the 'signal' voxels (i.e., the correlation r_{iM} between each image i and the median M, only in signal voxels). We calculated the standard deviation among noise voxels for each image S_i and the median S_M. Ultimately, the regression slope $Slope_i$ for each image (i) relative to the median image was determined by:

$$Slope_i = r_{iM} \frac{S_i}{S_M}$$

It is quite common for NeuroVault users to upload inverted contrasts and duplicates. For example, one might find two images representing the same contrast (such as House > Face) but with the signs reversed (i.e., Face > House). This creates problems for meta-analyses, as these images effectively cancel each other out when aggregated. Additionally, it is typical for users to upload multiple images of the same contrast, differing only by the covariate used in the group-level analysis. These can be considered duplicates, especially when the covariate does not influence the final estimate. To identify duplicates, we utilize the correlation matrix of the input samples. Image pairs with a correlation close to 1 are considered duplicates, while those with a correlation close to -1 are labeled as inverted contrasts. From the identified duplicates, we randomly selected one image from each pair. For pairs of inverted contrasts, we choose the image with a positive slope relative to the median image.

Finally, we removed images with extreme regression slope values relative to the population median using the interquartile range (IQR) method. First, we sorted the slopes in ascending order and calculated the first and third quartiles (Q1 and Q3). The IQR was defined as the difference between Q3 and Q1 (IQR = Q3 - Q1). Next, we determined the lower and upper bounds: the lower bound is calculated as Q1–1.5 * IQR, and the upper bound as Q3 + 1.5 * IQR. We compared each slope to these bounds, removing any images with slopes that were smaller than the lower bound or larger than the upper bound.

Manual selection
The manual meta-analysis served as an initial evaluation of the IBMA with NeuroVault data, as manually selected samples are less likely to include spurious or non-relevant images. Our primary focus was on the task description outlined in the method section of the paper, as well as the specific contrast of interest. To assist in our selection, we examined the image metadata in NeuroVault, specifically the image, file name, and contrast definition fields. For instance, in the case of a domain involving working memory tasks like the n-back task, we reviewed the paper associated with the collections containing images from this task. Images related to the Cognitive Atlas task were identified using the metadata field "cognitive_paradigm_cogatlas." We prioritized the section of the paper that describes the task used in the study and checked whether the contrast of interest for the meta-analysis (e.g., 2-back vs. baseline) was present. If the study did explore this contrast, we then examined all images available in the corresponding NeuroVault collection that met our preliminary and heuristic selection criteria. To locate the relevant images, we searched for the contrast of interest in various fields of the image metadata, including the Cognitive Atlas contrast ("cognitive_contrast_cogatlas"), image title, file name (under the "file" field in NeuroVault), and contrast definition (found in the "contrast_definition" field in NeuroVault).

Methodological consideration of IBMA using NeuroVault data

A preliminary evaluation of IBMA using NeuroVault data was performed on domains whose tasks are well represented in NeuroVault. We focused on working memory, motor, and emotion processing. For working memory, we used the working memory fMRI task paradigm and the n-back task. For the motor domain, we selected images linked to the motor fMRI task paradigm, the motor sequencing task, and the finger tapping task. Finally, for emotion processing, an emotion processing fMRI task paradigm. As the reference image for the three domains, we used effect size maps from the Human Connectome Project (HCP).

Stouffer's method is the most popular approach for combining individual images in IBMA[17]. This method combines test statistics, assuming that the input values are standardized to have a mean of zero and a variance of one (i.e., Z scores)[50]. However, best practices in meta-analysis suggest using values with meaningful units, incorporating uncertainty through standard errors instead of relying solely on Z or T statistics[3]. A significant challenge in neuroimaging, particularly with functional fMRI derivatives data, is that researchers typically share only T or Z statistic maps instead of actual estimates with their associated standard errors. Moreover, even if estimates were provided, we often lack information about the units of measurement for these estimates. For instance, the FSL fMRI pipeline scales mean brain intensity to 10,000[51], while the SPM pipeline targets a scale of 100[52], which usually aligns more closely with a value around 130. To address these concerns, we used standardized effect size as input for meta-analyses, recently proposed by Bossier, Nichols, and Moerkerke (2019)[53]. The use of standardized effect sizes, which have no units, it facilitates aggregating results from different studies.

We reconstructed the standardized effect size (i.e., Cohen's d) from the Z/T score maps using the sample size available in the image metadata in NeuroVault and assuming that a one-sample t-test was used. It is important to note that various types of analyses exist, such as two-sample analyses that compare different groups, correlations that assess the relationship between brain response and a covariate, and F-tests that compare two or more groups, among others. However, since most task-based fMRI studies utilize a one-sample t-test, this assumption is reasonable unless we have more specific information about each study.

By definition, the population standardized effect size for a one-sample analysis is the population mean divided by the population standard deviation ($d = \mu / \sigma$). For N subjects, we can thus compute an estimate of standardized effect sizes from a one-sample t-test as

$$\widehat{d} = \frac{t}{\sqrt{N}}$$

While the sample mean is unbiased for the population mean, the sample Cohen's d above is not unbiased for the population d. A bias correction due to Hedges is

$$\widehat{g} = h\,(N)\,\widehat{d}$$

where

$$h\,(N) = 1 - \frac{3}{4\,(N-1)-1}$$

In the following sections, we consider the effect estimate y_{kv} as the Hedges' g estimate:

$$y_{kv} = h\,(N_K)\,\widehat{d}_{kv}$$

Let y_{kv} be the effect estimate for contrast k at voxel v, $k = 1, \ldots, K$, $v = 1, \ldots, V$; denote the corresponding standard error be s_{kv}, and let N_k be the sample size.

Baseline estimator
Mean The baseline estimator used in this work is the mean, where input images are aggregated using the equation for each voxel:

$$\bar{y}_v = \frac{1}{K}\sum_k y_{kv}$$

Robust meta-analysis methods

Meta-analysis methods can be significantly affected by extreme observations. In most cases, except for random-effects meta-analysis, an extreme data point is treated as strong evidence of an effect and can skew the results towards significance. In a random-effects meta-analysis, however, an outlier can increase heterogeneity and ultimately reduce the significance of the findings. In this context, we proposed four robust alternative methods to the baseline mean. These methods are designed to tolerate a certain fraction of corrupted data while providing reasonable estimates of the overall effect. Although we will only define these methods for unit-based effects (such as Cohen's d or Hedge's g), they can also be applied analogously to test statistics.

Median The first robust method, the median, is denoted for a given voxel v as $\bar{y}_v^M$. The median has a "break-down point" of 50%, meaning it can remain unbiased even if up to 50% of the data is corrupted. Unlike the sample mean, the standard error of the median is affected by the actual distribution of the data. Therefore, the following result, unfortunately, relies on the assumption of normality:

$$Var\left(\bar{y}_v^M\right) = \frac{\pi}{2K} Var\left(y_v\right)$$

While concerns about outliers are valid, it is also essential to consider the variance of the data (denoted as $Var\left(y_v\right)$). In this regard, it is essential to recognize that the mean is generally more efficient than the median. When considering variance, the mean of K values has a variance that is 1/K times that of the original data. In contrast, the variance of the median is π/2, approximately 1.57 times larger. As a result, the standard deviation for the median is roughly 25% greater than the mean's.

Trimmed and winsorized mean The two other robust methods considered in this work are the Trimmed mean and the Winosorized mean. For these methods, we need additional notation: Let $y_{(1)v} \leq \cdots \leq y_{(k)v} \leq \cdots \leq y_{(K)v}$ be the *K* ordered effect sizes at voxel *v*. For either the trimmed or the winsorized mean, we make the decision to limit the influence of the (γ/2)100% smallest and (γ/2)100% largest observations for some γ < 0.5. In practice, since we always symmetrically drop data from both extremes, we only consider values γ that are multiples of 2/*K*. Denote $K\gamma = \gamma K/2$ as the number of observations to drop from each tail.

For the trimmed mean, we simply drop these observations and compute the mean of the remaining observations.

$$\bar{y}_v^T = \frac{1}{K - 2K_\gamma}\sum_{k=K_\gamma+1}^{K-K_\gamma} y_{(k)v}$$

For Winsorization, the extreme values are replaced with their nearest neighbors.

$$\bar{y}_v^W = \frac{1}{K}\left(K_\gamma\, y_{(K_\gamma+1)v} + \sum_{k=K_\gamma+1}^{K-K_\gamma} y_{(k)v} + K_\gamma\, y_{(K-K_\gamma)v}\right)$$

Regarding the choice of γ, we want to consider removing at least two studies, one from each tail, and so γ = 2/*K*. When there are many studies, there is no principled here one heuristic: If *K* = 20, then the minimum is γ = 2/20 = 10%, suggesting γ = 10% in general. Fortunately, for normally distributed data, the tails are so light that trimmed and winsorized mean with γ = 10% or even γ = 20% have good efficiency, though, with small *K*, it is probably best to stick to the lower value of 10%.

Weighted mean: fixed effect Hedges' g The weighted mean through a fixed effect meta-analysis is another robust estimator, where studies with smaller standard errors are weighted higher in the average of the input data. Using the equation for a fixed effect meta-analysis, assuming $Var\left(y_{kv}\right) = s_{kv}^2$, with the optimal weights $w_{kv} = s_{kv}^{-2}$, the fixed effect meta-analytic estimate is

$$\bar{y}_v^{FE} = \frac{1}{\Sigma_k w_k}\Sigma_k w_k y_{kv}$$

Where Bossier, Nichols, and Moerkerke (2019) defined the standard error of bias-corrected Cohen's d, i.e. Hedge's g, as

$$s_{kv} = \sqrt{\frac{(N-1)\left(1 + N\widehat{d}_{kv}\right)}{N(N-3)} h(N)^2 - \widehat{d}_{kv}^2}$$

Evaluating image-based meta-analyses

We assess the meta-analyses with all different combinations of parameters (i.e., image selection method and estimator approach) with the help of reference images from the task-fMRI group-average effect size maps from the HCP S1200 data release[40–44]. Specifically, we used the contrast of 2-back versus 0-back for working

memory. For the motor domain, we used the contrast representing the average of all motor movement blocks against the baseline in the HCP task. The contrast "Face vs. Shape" was our reference for emotion processing. Additional details regarding these tasks and their available contrasts can be found here[40]. HCP group-average maps were selected as reference standards for validation due to their coverage of distinct domains, scale, and data quality. The HCP is one of the most extensive and standardized neuroimaging datasets available, with $N = 1200$ participants completing identical protocols under rigorous quality control. This dataset has been effectively used as a reference for numerous studies. Specifically, we examined whether IBMA can detect robust, reproducible task effects observed in a large, homogeneous sample like the HCP maps, despite methodological heterogeneity, rather than expecting identical results. High correlations indicate successful identification of core task-related patterns; lower correlations may reflect either limitations of the IBMA approach or genuine heterogeneity in the literature. The level of correspondence informs us about both the quality of available data and the effectiveness of different image selection and aggregation strategies.

The comparisons between our IBMA results and the reference maps focused on evaluating image similarity and increased estimates in specific brain regions of interest. To quantitatively evaluate the similarity of the images, we calculated correlation coefficients between vertex-level unthresholded meta-analytic estimate maps from the IBMA and the reference unthresholded group-average effect size maps from the HCP. Since the reference maps were defined in CIFTI format, containing all grayordinates from the subcortical structure and cortical regions, we transformed the meta-analytic maps from the MNI152 space to the standard MNI fsLR 32 K 2-mm mesh surface space of the HCP, using the `mni152_to_fslr` function from the Neuromaps' transforms module[54]. For simplicity, we focused only on cortical regions for our evaluation. To assess the statistical significance of the spatial correlations between our IBMA results and the HCP reference maps, we performed spin permutation tests using 1000 rotations of the spherical projection of the HCP maps in surface space[55]. This approach preserves the spatial autocorrelation structure of the brain maps while generating a null distribution of correlation values under spatial independence. Subsequently, we quantitatively evaluated the increased estimated values in specific areas of interest across the domains. To achieve this, we focused our analysis on particular regions of interest, defined by selecting the top 10% of vertices for each reference map.

Step-by-Step framework for IBMA with NeuroVault

Performing image-based meta-analysis with NeuroVault data follows a systematic multi-step workflow that starts with data collection and moves through quality checks to final analysis. First, researchers download NeuroVault metadata tables and identify relevant cognitive domains using the Cognitive Atlas knowledge base. Then, they apply initial selection criteria to filter for fMRI-BOLD group-level statistical maps (T or Z scores) from studies with over ten subjects that are unthresholded, in MNI space, and have sufficient brain coverage (> 40%). Next, the selected images are downloaded from NeuroVault via their API, with each image assigned a unique identifier combining collection and image IDs. The statistical maps are then converted into a NiMARE Dataset object, which standardizes the data and allows converting T/Z scores to Cohen's d effect sizes using available sample size information. An important quality control step follows, involving automated heuristic filtering to remove spurious images (those with extreme Z values outside 1.96-50 range), duplicate maps, inverted contrasts, and non-statistical images identified by filename keywords like "ICA," "PCA," or "correlation." Finally, researchers set up an IBMAWorkflow in NiMARE with suitable parameters (estimator method, correction approach, diagnostic tools), run the meta-analysis, and produce detailed reports including corrected statistical maps, cluster tables, and jackknife diagnostics to evaluate how each study contributes to the overall results. For a hands-on guide to implementing this framework, see our interactive tutorial in Jupyter notebook, powered by MyBinder: https://github.com/neurostuff/2025-ohbm-ibma-neurovault.

Data availability

The imaging data for meta-analyses used in this project are publicly available for download at https://neurovault.org/.

Code availability

This project relied on multiple open-source Python packages, including: Jupyter[56], Matplotlib[57], Neuromaps[54], NiBabel[58], Nilearn[59], NiMARE[7,60], PyMARE[61], NumPy[62], Pandas[63], PtitPrince (*github.com/pog87/PtitPrince*), PySurfer[64], Scikit-learn[65], SciPy[66], Seaborn[67], and SurfPlot[68],. We also used the HCP software Connectome Workbench (wb_command version 1.5.0[69]. All code required to reproduce the analyses and figures in this paper is available on GitHub at https://github.com/NBCLab/large-scale-ibma. All data and resources that resulted from this paper (e.g., connectivity gradients and trained meta-analytic decoders) are openly disseminated and made available on the Open Science Framework (OSF) at https://osf.io/w7zcp/, including the links to the GitHub repository and figures.

Received: 6 April 2025; Accepted: 15 September 2025
Published online: 21 October 2025

References

1. Kent, J. et al. Neurosynth compose: A Web-Based platform for flexible and reproducible neuroimaging Meta-Analysis. *OSF* https://doi.org/10.31219/osf.io/ywxu7_v1 (2025).
2. Lazar, N. A., Luna, B., Sweeney, J. A. & Eddy, W. F. Combining brains: A survey of methods for statistical pooling of information. *NeuroImage* **16** (2), 538–550. https://doi.org/10.1006/nimg.2002.1107 (2002).

3. Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. & Nichols, T. E. Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage* **45** (3), 810–823. https://doi.org/10.1016/j.neuroimage.2008.12.039 (2009).
4. Gorgolewski, K. J. et al. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinformatics*. **9**, 8. https://doi.org/10.3389/fninf.2015.00008 (2015).
5. Hammond, C. J. et al. A Meta-Analysis of fMRI studies of youth cannabis use: alterations in executive Control, social Cognition/Emotion processing, and reward processing in cannabis using youth. *Brain Sci.* **12** (10), 1281. https://doi.org/10.3390/brainsci12101281 (2022).
6. Menuet, R., Meudec, R., Dockès, J., Varoquaux, G. & Thirion, B. Comprehensive decoding mental processes from Web repositories of functional brain images, *Sci. Rep.* **12**(1), 1 https://doi.org/10.1038/s41598-022-10710-1 (2022).
7. Salo, T. et al. NiMARE: neuroimaging Meta-Analysis research environment. *Aperture Neuro*. **3**, 1–32. https://doi.org/10.52294/001c.87681 (2023).
8. Costafreda, S. G. Meta-Analysis, Mega-Analysis, and Task Analysis in fMRI Research, *Philosophy, Psychiatry, & Psychology* **18**(4), 275–277, (2011).
9. Costafreda, S. G. Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. *Front. Neuroinform*. **3** https://doi.org/10.3389/neuro.11.033.2009 (2009).
10. Norman, L. J. & Shaw, P. Harnessing mega-analysis in the era of 'big data' neuroimaging. *Neuropsychopharmacology* **50** (1), 332–334. https://doi.org/10.1038/s41386-024-01964-6 (2025).
11. Steele, N. et al. Image-Based Meta- and Mega-Analysis (IBMMA): A unified framework for Large-Scale, Multi-Site, neuroimaging data analysis. *BioRxiv* https://doi.org/10.1101/2025.06.16.657725 (2025).
12. Zugman, A. et al. Mega-analysis methods in ENIGMA: the experience of the generalized anxiety disorder working group. *Hum. Brain. Mapp.* **43** (1), 255–277. https://doi.org/10.1002/hbm.25096 (2022).
13. Poldrack, R. A. & Gorgolewski, K. J. Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* **17** (11), 1510–1517. https://doi.org/10.1038/nn.3818 (2014).
14. Poline, J. B. et al. Data sharing in neuroimaging research. *Front. Neuroinform*. **6** https://doi.org/10.3389/fninf.2012.00009 (2012).
15. White, T., Blok, E. & Calhoun, V. D. Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Hum. Brain. Mapp.* **43** (1), 278–291. https://doi.org/10.1002/hbm.25120 (2022).
16. Fisher, R. A. Statistical Method For Research Workers, (1934).
17. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams, R. M. Jr *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1*. in The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1. Oxford, England: Princeton Univ. Press, pp. xii, 599. (1949).
18. Fox, P. T. et al. Functional volumes modeling: scaling for group size in averaged images. *Hum. Brain. Mapp.* **8**, 2–3. https://doi.org/10.1002/(SICI)1097-0193(1999)8:2/3%3C143::AID-HBM12%3E3.0.CO;2-9 (1999).
19. Fox, P. T., Parsons, L. M. & Lancaster, J. L. Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* **8** (2), 178–187. https://doi.org/10.1016/S0959-4388(98)80138-4 (1998).
20. Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J. H. & Zamarripa, F. Functional volumes modeling: theory and preliminary assessment. *Hum. Brain. Mapp.* **5** (4), 306–311. https://doi.org/10.1002/(SICI)1097-0193(1997)5:4%3C306::AID-HBM17%3E3.0.CO;2-B (1997).
21. Nielsen, F. Å. & Hansen, L. K. Modeling of activation data in the BrainMap™ database: detection of outliers. *Hum. Brain. Mapp.* **15** (3), 146–156. https://doi.org/10.1002/hbm.10012 (2002).
22. Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F. & Fox, P. T. Activation likelihood Estimation meta-analysis revisited. *NeuroImage* **59** (3), 2349–2361. https://doi.org/10.1016/j.neuroimage.2011.09.017 (2012).
23. Wager, T. D., Jonides, J. & Reading, S. Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage* **22** (4), 1679–1693. https://doi.org/10.1016/j.neuroimage.2004.03.052 (2004).
24. Wager, T. D., Phan, K. L., Liberzon, I. & Taylor, S. F. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *NeuroImage* **19** (3), 513–531. https://doi.org/10.1016/S1053-8119(03)00078-8 (2003).
25. Wager, T. D., Lindquist, M. & Kaplan, L. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cognit. Affect. Neurosci.* **2** (2), 150–158. https://doi.org/10.1093/scan/nsm015 (2007).
26. Ma, H. et al. Abnormal amygdala functional connectivity and deep learning classification in multifrequency bands in autism spectrum disorder: A multisite functional magnetic resonance imaging study. *Hum. Brain. Mapp.* **44** (3), 1094–1104. https://doi.org/10.1002/hbm.26141 (2023).
27. Cui, Y. et al. Consistent brain structural abnormalities and multisite individualised classification of schizophrenia using deep neural networks. *Br. J. Psychiatry*. **221** (6), 732–739. https://doi.org/10.1192/bjp.2022.22 (2022).
28. Fiorito, A. M. et al. Are brain responses to emotion a reliable endophenotype of schizophrenia? An Image-Based functional magnetic resonance imaging Meta-analysis. *Biol. Psychiatry*. **93** (2), 167–177. https://doi.org/10.1016/j.biopsych.2022.06.013 (2023).
29. Hellewell, S. C., Nguyen, V. P. B., Jayasena, R. N., Welton, T. & Grieve, S. M. Characteristic patterns of white matter tract injury in sport-related concussion: an image based meta-analysis. *NeuroImage: Clin.* **26**, 102253. https://doi.org/10.1016/j.nicl.2020.102253 (2020).
30. Lamm, C., Decety, J. & Singer, T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* **54** (3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014 (2011).
31. Luijten, M., Schellekens, A. F., Kühn, S., Machielse, M. W. J. & Sescousse, G. Disruption of reward processing in addiction : an Image-Based Meta-analysis of functional magnetic resonance imaging studies. *JAMA Psychiatry*. **74** (4), 387–398. https://doi.org/10.1001/jamapsychiatry.2016.3084 (2017).
32. Lukow, P. B. et al. Neural correlates of emotional processing in psychosis risk and onset – A systematic review and meta-analysis of fMRI studies. *Neurosci. Biobehavioral Reviews*. **128**, 780–788. https://doi.org/10.1016/j.neubiorev.2021.03.010 (2021).
33. Schulze, L., Schulze, A., Renneberg, B., Schmahl, C. & Niedtfeld, I. Neural correlates of affective disturbances: A comparative Meta-analysis of negative affect processing in borderline personality disorder, major depressive disorder, and posttraumatic stress disorder. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*. **4** (3), 220–232. https://doi.org/10.1016/j.bpsc.2018.11.004 (2019).
34. Schulze, L., Schmahl, C. & Niedtfeld, I. Neural correlates of disturbed emotion processing in borderline personality disorder: A multimodal Meta-Analysis. *Biol. Psychiatry*. **79** (2), 97–106. https://doi.org/10.1016/j.biopsych.2015.03.027 (2016).
35. Witt, S. T., van Ettinger-Veenstra, H., Salo, T., Riedel, M. C. & Laird, A. R. What executive function network is that? An Image-Based Meta-Analysis of network labels. *Brain Topogr.* **34** (5), 598–607. https://doi.org/10.1007/s10548-021-00847-z (2021).
36. Laird, A. R., Lancaster, J. L. & Fox, P. T. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* **3** (1), 65–78. https://doi.org/10.1385/ni:3:1:065 (2005).
37. Fox, P. T. et al. BrainMap taxonomy of experimental design: description and evaluation. *Hum. Brain Mapp.* **25** (1), 185–198. https://doi.org/10.1002/hbm.20141 (2005).
38. Laird, A. R. et al. ALE Meta-Analysis workflows via the brainmap database: progress towards A probabilistic functional brain atlas. *Front. Neuroinform*. **3, p.** (23). https://doi.org/10.3389/neuro.11.023.2009 (2009).
39. Laird, A. R. et al. The brainmap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res. Notes*. **4** (1), 349. https://doi.org/10.1186/1756-0500-4-349 (2011).

40. Barch, D. M. et al. Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033 (2013).
41. Smith, S. M. et al. Resting-state fMRI in the human connectome project. *NeuroImage* **80**, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039 (2013).
42. Uğurbil, K. et al. Pushing Spatial and Temporal resolution for functional and diffusion MRI in the human connectome project. *NeuroImage* **80**, 80–104. https://doi.org/10.1016/j.neuroimage.2013.05.012 (2013).
43. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *NeuroImage* **80**, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041 (2013).
44. Van Essen, D. C. et al. The human connectome project: A data acquisition perspective. *NeuroImage* **62** (4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018 (2012).
45. Poldrack, R. A. et al. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinform.* **5**, 17. https://doi.org/10.3389/fninf.2011.00017 (2011).
46. Dockès, J., Oudyk, K., Torabi, M., de la Vega, A. I. & Poline, J. B. Mining the neuroimaging literature. *eLife* **13** https://doi.org/10.7554/eLife.94909.1 (2024).
47. de la Vega, A. et al. Neuroscout, a unified platform for generalizable and reproducible fMRI research. *eLife* **11**, e79277. https://doi.org/10.7554/eLife.79277 (2022).
48. Lefort-Besnard, J., Nichols, T. E. & Maumet, C. Same Data Meta Analysis for Neurimaging Multiverse Data, presented at the INCF Neuroinformatics Assembly 2024. Accessed: Jan. 09, 2025. [Online]. Available: https://inria.hal.science/hal-04697056 (2024).
49. Maumet, C. & Nichols, T. E. Minimal Data Needed for Valid & Accurate Image-Based fMRI Meta-Analysis, *bioRxiv*. (2016). https://doi.org/10.1101/048249
50. Camille, M. & Thomas, N. An SPM toolbox for neuroimaging Image-Based Meta-Analysis. *Front. Neuroinform.* **8** https://doi.org/10.3389/conf.fninf.2014.18.00025 (2014).
51. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62**(2), 782–790. (2012). https://doi.org/10.1016/j.neuroimage.2011.09.015
52. Friston, K., Ashburner, J., Kiebel, S., Nichols, T. & Penny, W. (eds) *Statistical Parametric Mapping* (Academic, 2007). https://doi.org/10.1016/B978-0-12-372560-8.50052-8
53. Bossier, H., Nichols, T. E. & Moerkerke, B. Standardized effect sizes and image-based meta-analytical approaches for fMRI data. *bioRxiv*. (2019). https://doi.org/10.1101/865881
54. Markello, R. D. et al. neuromaps: structural and functional interpretation of brain maps. *bioRxiv*. (2022). https://doi.org/10.1101/2022.01.06.475081
55. Alexander-Bloch, A. F. et al. On testing for Spatial correspondence between maps of human brain structure and function. *NeuroImage* **178**, 540–551. https://doi.org/10.1016/j.neuroimage.2018.05.070 (2018).
56. Kluyver, T. et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. *Position. Power Acad. Publishing: Players Agents Agendas*. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87 (2016).
57. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9** (03), 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).
58. Brett, M. et al. *nipy/nibabel: 3.2.1*. Zenodo. (2020). https://doi.org/10.5281/zenodo.4295521
59. Abraham, A. et al. Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8** https://doi.org/10.3389/fninf.2014.00014 (2014).
60. Salo, T. et al. *neurostuff/NiMARE: 0.4.1*. Zenodo. (2024). https://doi.org/10.5281/zenodo.14183422
61. Yarkoni, T., Salo, T., Peraza, J. A. & Nichols, T. E. *neurostuff/PyMARE: 0.0.8* Zenodo. (2024). https://doi.org/10.5281/zenodo.13743687
62. van der Walt, S., Colbert, S. C. & Varoquaux, G. The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13** (2), 22–30. https://doi.org/10.1109/MCSE.2011.37 (2011).
63. McKinney, W. Data Structures for Statistical Computing in Python, presented at the Python in Science Conference, Austin, Texas, pp. 56–61. (2010). https://doi.org/10.25080/Majora-92bf1922-00a
64. Waskom, M. et al. *nipy/PySurfer: 0.11.0*. Zenodo. (2020). https://doi.org/10.5281/zenodo.3905195
65. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 6 (2011).
66. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*. 1–12. https://doi.org/10.1038/s41592-019-0686-2 (2020).
67. Waskom, M. L. Seaborn: statistical data visualization. *J. Open. Source Softw.* **6**, 3021. https://doi.org/10.21105/joss.03021 (2021).
68. Gale, D. J., Vos de Wael, R., Benkarim, O. & Bernhardt, B. Surfplot: Publication-ready brain surface figures. *Oct* **14** https://doi.org/10.5281/zenodo.5567926 (2021). Zenodo.
69. Marcus, D. et al. Informatics and data mining tools and strategies for the human connectome project. *Front. Neuroinform.* **5** https://doi.org/10.3389/fninf.2011.00004 (2011).

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions during the revision of this manuscript. Special thanks to the FIU Instructional & Research Computing Center (IRCC, http://ircc.fiu.edu) for providing the HPC and computing resources that contributed to the research results reported in this paper.

Author contributions

ARL, JAP, AdlV, JBP, TEN, and JDK conceived and designed the project. JAP, AdlV, RWB, and JDK analyzed data. JAP, RWB, and JDK contributed scripts and pipelines. JAP, TEN, and ARL wrote the paper, and all authors contributed to the revisions and approved the final version.

Funding

Funding for this project was provided by NIH R01-MH096906.

Declarations

Competing interests

The authors declare no competing interests.

Ethical statement

The Human Connectome Project provided the ethics and consent needed for the study and dissemination

of HCP data. This secondary data analysis was approved by the Institutional Review Board of Florida International University.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-20328-8.

Correspondence and requests for materials should be addressed to J.A.P. or A.R.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.