

# UC San Diego

## UC San Diego Previously Published Works

### Title

Quality over quantity: powering neuroimaging samples in psychiatry

### Permalink

<https://escholarship.org/uc/item/0vv9c0zj>

### Journal

Neuropsychopharmacology, 50(1)

### ISSN

0893-133X

### Authors

Makowski, Carolina

Nichols, Thomas E

Dale, Anders M

### Publication Date

2024-11-01

### DOI

10.1038/s41386-024-01893-4

Peer reviewed



## REVIEW ARTICLE

## Quality over quantity: powering neuroimaging samples in psychiatry

Carolina Makowski <sup>1</sup>✉, Thomas E. Nichols <sup>2</sup> and Anders M. Dale<sup>3</sup>

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

Neuroimaging has been widely adopted in psychiatric research, with hopes that these non-invasive methods will provide important clues to the underpinnings and prediction of various mental health symptoms and outcomes. However, the translational impact of neuroimaging has not yet reached its promise, despite the plethora of computational methods, tools, and datasets at our disposal. Some have lamented that too many psychiatric neuroimaging studies have been underpowered with respect to sample size. In this review, we encourage this discourse to shift from a focus on sheer increases in sample size to more thoughtful choices surrounding experimental study designs. We propose considerations at multiple decision points throughout the study design, data modeling and analysis process that may help researchers working in psychiatric neuroimaging boost power for their research questions of interest without necessarily increasing sample size. We also provide suggestions for leveraging multiple datasets to inform each other and strengthen our confidence in the generalization of findings to both population-level and clinical samples. Through a greater emphasis on improving the quality of brain-based and clinical measures rather than merely quantity, meaningful and potentially translational clinical associations with neuroimaging measures can be achieved with more modest sample sizes in psychiatry.

*Neuropsychopharmacology* (2025) 50:58–66; <https://doi.org/10.1038/s41386-024-01893-4>

## INTRODUCTION

Psychiatry has greatly benefitted from the integration of neuroimaging techniques to elucidate the neuroanatomical and functional complexity of disorders of the mind. However a flood of concerns regarding effect sizes, alongside recommendations of sample sizes in the thousands [1, 2] to reliably measure brain-behavior associations, has caught psychiatric imaging between a rock and a hard place. Collecting thousands of individuals is prohibitively difficult for many research questions in psychiatry, layered with additional challenges that accompany the recruitment and retention of patients. Many researchers have turned towards existing large consortia-level datasets, although this comes at the cost of increased heterogeneity and potential loss of granularity in understanding individual differences, which may cloud key pursuits in precision medicine. It is also unlikely that merely increasing sample sizes will propel psychiatric neuroimaging towards translational models with clinical utility. Is it really the case that overcoming stagnancy of neuroimaging methods in translational psychiatry boils down to the need for larger samples? Or is a shift in focus from quantity to *quality* of collected behavioral and neuroimaging data needed? This review will emphasize the latter, and walk through critical decision points in statistical modeling and phenotype selection to boost effect sizes between brain-derived measures and psychiatric outcomes, even with more modest sample sizes. We capture these decision points across three key stages of investigation, beginning with (1) experimental paradigm and study design choices in prospective

studies; and considerations that can be applied to both prospective and retrospective studies involving (2) data cleaning and modeling; and (3) analyses. We end with a discussion of how multiple datasets can be leveraged in predicting key outcomes in psychiatry. A summary of the considerations we discuss in this article can be found in Fig. 1.

## CONSIDERATION 1: EXPERIMENTAL PARADIGM AND STUDY DESIGN CHOICES

## Sample size considerations

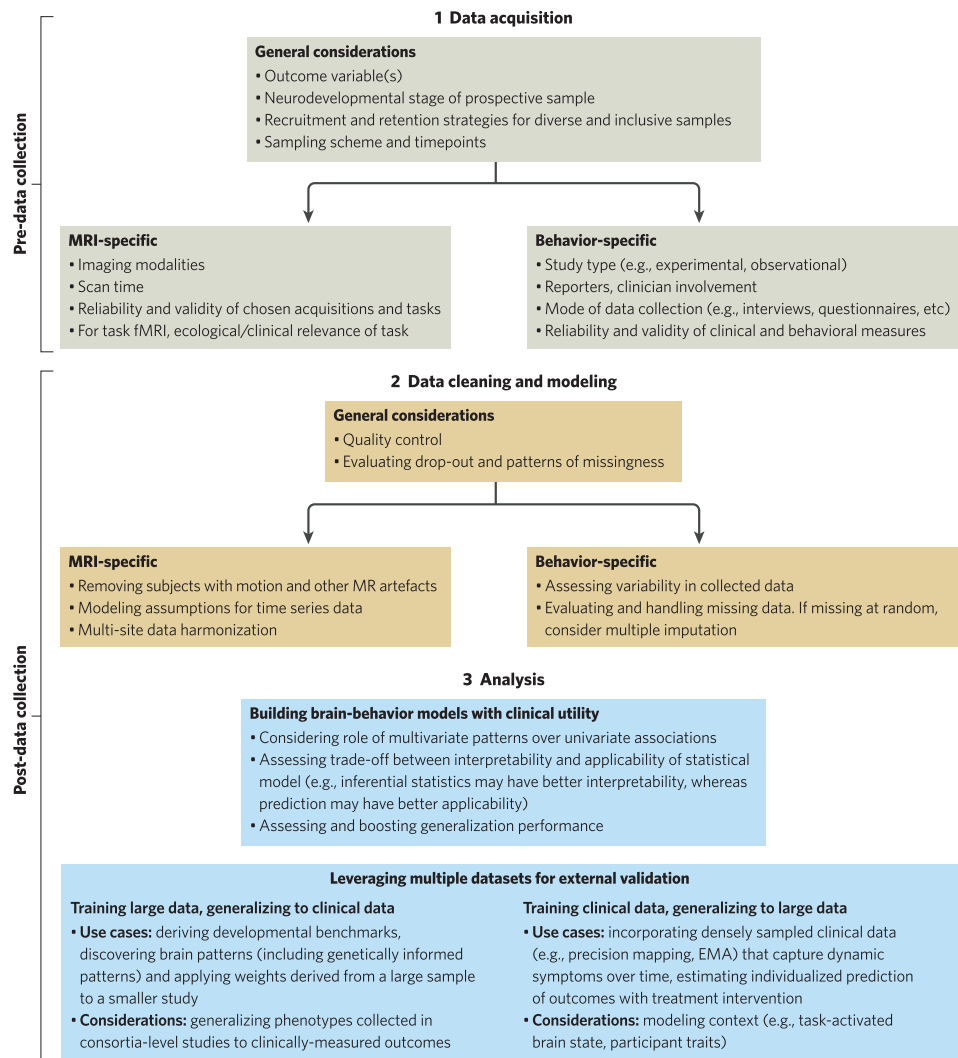
Although we will emphasize “quality” of prospective and collected data in the proceeding sections, we still acknowledge the importance of “quantity” or sample size estimates themselves, and the power required to detect an effect of interest. Statistical power is simply the probability of detecting a true (non-null) effect. Small sample sizes have low power and provide weak evidence when an effect is detected [3]. In reality, many neuroimaging studies have been underpowered [4], which in turn, increases the probability that purported significant findings were in reality false. This is particularly concerning given that many underpowered neuroimaging studies with potentially inflated type I (false positive) errors have been highly cited [5], facilitating the pursuit and propagation of misleading research directions.

Although researchers have adopted practices for controlling type I errors, such as raising the significance threshold, this comes

<sup>1</sup>Department of Radiology, University of California San Diego, San Diego, CA, USA. <sup>2</sup>Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. <sup>3</sup>Departments of Radiology and Neurosciences, University of California San Diego, San Diego, CA, USA. ✉email: cmakowski@ucsd.edu

Received: 22 March 2024 Revised: 6 May 2024 Accepted: 24 May 2024

Published online: 20 June 2024



**Fig. 1** Considerations discussed throughout the review to improve power in detecting brain-behavior associations, beyond increasing sample size. EMA Ecological Momentary Assessment.

at the cost of power, with an increased rate of type II (false negative) errors. Once data has been collected, there is not much that can be done to modify the sample size, but there are still steps that researchers working in psychiatric neuroimaging can take to help mitigate type II errors and magnify the effect sizes of meaningful brain-behavior associations, which will be discussed throughout the rest of this review.

### Choices in neuroimaging acquisition

Magnetic Resonance Imaging (MRI) has become increasingly cost-effective, but still involves a high load of monetary and personnel resources that warrants strong justification. It can feel enticing to incorporate as many imaging modalities as possible into a single brain MRI scan protocol. However, prioritizing imaging acquisitions relevant for an investigator's core question of interest (e.g., structural MRI for mapping developmental trajectories of brain morphology) may free up additional resources to recruit more participants. Rs-fMRI in particular has gained notable popularity over the years, in part due to ease of acquisition and comparison across multiple studies compared to task-based fMRI. Although the brain at rest or in the absence of goal-directed tasks may be relevant for specific questions, the explosion of rs-fMRI in human brain mapping efforts may also be influencing investigators to incorporate such protocols despite appropriate rationale for their

inclusion. Recent studies and commentaries have emphasized a need to shift our focus back to task fMRI [6–8]. There is strong evidence suggesting that task fMRI typically yields stronger predictions of behavior, particularly cognition, compared to resting state functional connectivity [9–14]. This points to the importance and utility of engaging participants' attention with a directed task in cognitive neuroimaging study design [15, 16]. Although we focus predominantly on structural and task fMRI throughout this review, we are enthusiastic about promising strides in the development of ecologically relevant paradigms through portable functional imaging technology (e.g., functional near-infrared spectroscopy [fNIRS], electroencephalography [EEG], magnetoencephalography [MEG]).

The intentional integration of task-based imaging paradigms alongside task-relevant behavior may be a clear path forward to boost power for effects of interest in psychiatry with smaller samples. However, many commonly used tasks (e.g., n-back for working memory, Stroop for executive function, monetary incentive tasks for reward/motivation, etc.), which tend to be implemented in large consortia studies, may not be enriched for associations with psychopathology per se [17]. Overcoming these shortcomings will require the creation and integration of novel task paradigms that target valid cognitive constructs associated with psychiatric disorders [7, 18]. It will be important for funding

agencies and institutions to support such initiatives in smaller samples before being substantiated in future population-based studies.

### Choices in behavioral phenotypes

With rebounding reproducibility concerns of studies associating brain measures with behavior, neuroimaging is often the main target of scrutiny. For example, a press release of Marek & Tervo-Clemmens et al. [1] influential paper was featured in the New York Times, declaring “*Brain-Imaging studies hampered by small data sets*” [19]. However, the reliability and validity of chosen behavioral outcomes is equally as important. There are a wide array of decision points to reflect upon when choosing behavioral measures to incorporate in data collection. These decision points non-exhaustively include: study type (e.g., experimental vs observational), reporter (e.g., self-report, caregiver, teacher, clinician, etc.), and mode of collection (structured and semi-structured interviews, questionnaires, digital tasks), which should be accompanied by an assessment of both reliability and validity (e.g., internal, external, ecological) of the chosen behavioral paradigm. The degree of subjectivity in a particular measure can also introduce additional noise in behavioral measures, which in turn, reduces reliability [20]. There has been a recent emphasis to move towards “precision behavioral mapping” [21], underscoring the importance of focusing on behavioral measures to power associations between biology and observable psychiatric traits. Increasing the amount of between-subject variability in measured behavioral phenotypes will also help boost effect sizes, without necessarily increasing sample size [22].

The pursuit of uncovering neuroimaging-based biomarkers of psychiatric disorders necessitates an acknowledgment of the limitations of the diagnostic framework that psychiatry operates within. The symptoms and syndromes defining a particular psychiatric disorder have been established independently of any known biological constructs, oftentimes yielding very small effect sizes in psychiatric case-control differences in MRI measures. Two individuals could meet criteria for the same diagnosis but have an entirely non-overlapping set of symptoms. In turn, such heterogeneity in diagnostic categories may preclude discovery of clinically translational neuroimaging differences. For example, many studies have yielded inconsistent, even null findings, in longitudinal structural MRI-based markers of patients with schizophrenia spectrum disorders [23–25]. In one sample of psychosis patients early in the course of their illness, group differences in cortical and subcortical trajectories were only found when the sample was stratified by the presence of symptoms that are known to influence functional outcomes [26, 27]. Results such as these emphasize the importance of focusing on clinically meaningful symptom clusters, rather than traditional case-control designs, to improve translational efforts of neuroimaging studies in psychiatry. On the flip side, it is entirely plausible to have two patients with similar symptom profiles, but falling into different diagnostic categories. For example, it is not uncommon to see overlapping symptoms of social withdrawal and expressivity deficits in patients with autism and schizophrenia, warranting transdiagnostic or dimensional approaches of brain-behavior associations with symptoms. It has been encouraging to see such transdiagnostic efforts in recent years, although this may necessitate the integration of larger sample sizes, depending on the analytical framework (see consideration 3).

### Longitudinal sampling

Longitudinal designs naturally offer an additional boost in power compared to cross-sectional studies. With the addition of one timepoint per subject, associations of gray matter volumes with age were boosted by 35% in a recent investigation [22]. Choosing brain measures informed by the neurodevelopmental stage of the target sample should also be considered. For instance, MRI-

derived measures that reach their maturational peak earlier in development (e.g., subcortical volumes around puberty [28]; cortical surface area in late childhood [29]) may have lower between-subject variability in adolescent/adult patient samples, and thus may yield smaller effect sizes in brain-behavior measures, particularly if working in a longitudinal framework.

Dense sampling across both neuroimaging and behavioral measures, where a large number of longitudinal datapoints are collected per individual, can also help boost discovery of meaningful brain-behavior associations [30–32]. In this vein, the above-mentioned concept of “precision behavioral mapping” can be extended to neuroimaging data as well. Indeed, functional MRI has seen many examples of such precision mapping with longer scan times and repeated samplings over short intervals of time [33–36]. It has been suggested that scan time may be just as important as sample size when designing a study [37]. Dense sampling can provide rich information to map onto more dynamic symptom fluctuations and other within-subject changes that group-level statistical approaches would otherwise miss. This framework will be revisited later in this review, when we highlight recent research and future avenues for integrating densely sampled neuroimaging and behavioral data into precision psychiatry efforts.

Thus far, we have highlighted recommendations that apply solely to prospective study designs. Choosing the right imaging modalities, phenotypes, and measures is incredibly important, but the decision points do not stop there. There are many steps that can still be taken to potentially boost brain-behavior effect sizes with existing data. The remaining text will focus on considerations that can be applied to both prospective and retrospective study designs, across both data modeling and analysis frameworks.

## CONSIDERATION 2: DATA CLEANING AND MODELING

### Considerations for inclusion and modeling of neuroimaging data

Neuroimaging studies routinely require a description of quality assurance protocols to ensure the inclusion of good quality and low motion scans in analysis. This is particularly germane for psychiatric disorders, as head motion may be increased in certain patient populations, for instance in conditions characterized by motor symptoms or impulsivity [38, 39]. The inclusion of scans with motion artifacts can yield different effect sizes and results for interpretation compared to a quality-controlled subset of the same sample. Longitudinal trajectories of structural [40–42] and functional [43] neuroimaging markers can also be greatly impacted by motion, which holds implications for psychiatric neuroimaging studies investigating developmental markers of psychopathology. Readers are referred to previous work outlining recommendations for handling motion in psychiatric neuroimaging studies [44].

Throughout these decision points, it is important to keep in mind any assumptions that are being made in data modeling. Commonly adopted methods in the field may not always reflect the most biologically plausible model. For instance, the majority of studies using task-based fMRI time series data assume a single fixed impulse response function [45, 46]. In reality, the task activation signals captured by fMRI are much more complex and exhibit variability in peak timing and stimuli-related dynamics [47–50]. The spatiotemporal variability of the hemodynamic response (e.g., across brain regions and task conditions) may thus be better captured by empirical modeling of fMRI time series data [47–50]. It is worth noting, however, that such modeling frameworks do entail more degrees of freedom in their estimation, which may be more suitable for larger datasets that can afford the consequential dip in statistical power. Deciding whether or not an estimation of the hemodynamic response function is necessary certainly depends on the sample size and research question at

hand (e.g., it may be harder to detect group differences, but could boost power for brain-behavior predictions [51]), and illustrates how modeling decisions can have important implications for the measurement of brain-behavior associations and observed power.

Data harmonization efforts to compare neuroimaging data across existing studies with different scanning parameters can also help boost power. Although beyond the scope of this review, readers are encouraged to refer to ongoing efforts for harmonization of structural [52], diffusion [53], and functional MRI [54].

### Handling missing behavioral data

Imputation methods for rescuing missing behavioral data have been a valuable tool to retain a larger number of participants in analysis, especially in studies with lengthy protocols and/or longitudinal designs, where it becomes increasingly inevitable that some data points will not be collected. Missing data cannot always be reliably recovered, but if the missing data is correlated with observed data, imputation could be a viable option. Otherwise, researchers may want to consider a complete case analysis to enhance sensitivity of results, despite a likely drop in sample size. Note that power calculations must consider the anticipated dropout rate, and it is recommended to inflate recruitment numbers to ensure the desired sample size will be available at the end of the study.

Understanding patterns of missingness in a given dataset can be critical to ensure your results are in fact generalizable to the sample you are hoping to describe, and can help identify directions to improve upon in subsequent data collection efforts. Imputation methods may vary depending on whether data are missing completely at random (i.e., missing data values are distributed randomly over the sample and not related to any study outcomes), missing at random (i.e., probability of a value being missing is unrelated to its value after covarying for other variables), or exhibit non-random patterns of missingness (i.e., missing data is associated with certain values of a measured outcome or covariate) [55–58]. For instance, Gard and colleagues demonstrated an increased rate of missing data in the Adolescent Brain Cognitive Development<sup>SM</sup> Study (ABCD Study<sup>®</sup>) in participants with lower socioeconomic resources and in youth who identify as racial and/or ethnic minorities [59]. This pattern of disproportionate missingness in marginalized groups was particularly amplified in functional MRI data from the ABCD Study database that passed recommended quality control flags. Given the probability sampling design of the ABCD Study and large participant pool, it is possible for researchers working with ABCD Study data and similarly designed datasets to recover socio-demographic patterns in the data that match the target population sample and in turn, increase generalizability of findings. However, this is not to be interpreted as a replacement for critical recruitment strategies that need to be prioritized to ensure diverse representation of participants included in our analytic samples. This includes incorporating culturally-informed community outreach and recruitment strategies to build trust with patients from underrepresented minority groups to engage in research [59, 60].

In cases where data is missing at random, multiple imputation (e.g., through the Multivariate Imputation by Chained Equations [MICE] [61] procedure) can be considered to replace missing values. This involves replacement of missing values through an imputation regression model that predicts (with some intentionally added noise) missing values based on the missing variable's relationship with other variables in the dataset over several iterations. The resultant imputed datasets are then analyzed and pooled together into a single imputed estimate, with estimated uncertainty around the missing data. Readers are encouraged to consult [55] and [58] for recommendations to best handle missing data in clinical datasets.

### CONSIDERATION 3: ANALYTICAL FRAMEWORK

Brain-behavior associations are often shrouded by the high dimensionality of data for both imaging (i.e., across space for structural/diffusion MRI, and space and time for fMRI) and behavioral domains, which can be particularly challenging for brain-behavior mapping in smaller psychiatric samples. Below are considerations for multivariate analytic frameworks that appropriately handle the high dimensional nature of data underlying psychiatric imaging objectives, as well as acknowledging their limitations and future directions for improvements.

Univariate associations have dominated the field of psychiatric neuroimaging research, but the distributed sparse nature of many brain-behavior associations warrants a multivariate approach, which may afford the possibility to utilize smaller samples [8, 62–64]. Symptoms of psychiatric disorders do not 'reside' within a single brain region [65]. Region-of-interest-based approaches have also been popular in the field of psychiatric neuroimaging, but this selection of specific features gives a false impression that we already have a solid understanding of the neurobiological underpinnings of psychiatric disorders, which unfortunately is not the current state of the field (Abdallah et al. [66]). Further, the literature that motivates the selection of these regions may be biased by underpowered samples and findings that are not reproducible, alongside a publishing infrastructure that has historically incentivized the reporting of non-null results. This is not to say that theory-driven approaches of specific brain regions and circuits should be completely avoided. In these cases, researchers are encouraged to clearly articulate and justify narrowing of their search space to pre-defined regions of interest, and to pre-register their hypotheses.

Instead, data-driven multivariate analyses may offer a more biologically plausible framework to integrate the known complexity of brain circuitry and networks involved in clinical constructs. Multivariate modeling also bypasses the need for multiple comparison correction inherent to mass univariate analyses, and oftentimes yields larger effect sizes [8, 67]. This, in turn, may facilitate the inclusion of smaller sample sizes in psychiatry in particular use cases.

As with all models, multivariate models and machine learning algorithms still come with their own set of challenges. These include data leakage from training to test sets in cross-validation, risks of overfitting, and sample variability [68, 69]. Recent estimates have suggested that there are instances where multivariate prediction can be carried out with 50+ subjects in the discovery sample [8, 70], particularly for working memory-related functional activation and general cognitive ability [8]. Overfitting and inflation of predictive accuracy is notably increased in smaller samples [71–73]. In a review of 118 neuroimaging studies of pattern recognition applications to five psychiatric disorders, overfitting was a pervasive issue for studies with fewer than 50 participants [74]. Imbalanced sample sizes are also common in psychiatric neuroimaging studies, particularly for less frequent disorders or for prediction of rare outcomes. This can lead to inflated prediction accuracy, given that the classifier will naturally be more successful in predicting labels from the majority class. Incorporation of balanced accuracy metrics and/or budgeting resources to oversample rare cases can help mitigate such sample imbalances [75].

There have been concerted efforts to push psychiatry and related fields towards using predictive models with machine learning, rather than relying simply on explanation [76]. Others have lamented that although prediction is promising and likely necessary for clinical translation, many existing methods yield results that lack generalizability, with poor performance in truly independent datasets [77]. There are exceptions to this (see [78]), but generalizing to external datasets in psychiatry has been notoriously difficult. However, improving power in the test dataset for successful out-of-sample generalization will be critical if the



goal is to implement predictive models in the clinic. Recent work in four large openly available datasets has suggested that internal (within-dataset) prediction performance is typically within  $r \sim 0.2$  of external (cross-dataset) performance, which may help researchers gain a sense of the sample size that might be required to power their generalization dataset [79].

The differences in measured behavioral or clinical outcomes in independent train and test datasets will contribute to the above-mentioned lack of generalizability in psychiatric prediction. Given these challenges, is it feasible for psychiatric neuroimaging to transition to such predictive frameworks? Or are there cases where more classical inferential statistics are beneficial to understand the variance explained in observable behavior by brain-based measures?

The answer largely lies in a tradeoff between prediction and interpretability [20, 80]. Interpretable models allow us to better understand the cognitive and neurobiological mechanisms of a particular outcome. Simpler statistical models may be more beneficial in this case. By contrast, a highly accurate predictive model may not yield any meaningful insight into neurobiology (i.e., the model operates within a 'black box'), but simply points to a list of features that are highly predictive and could be adopted in real-world applications. If the goal is to show the utility of neuroimaging in clinical practice, it will be important for such models to offer both predictive power and neurobiological insight [20]. There is a place for both machine learning and simpler statistical models in psychiatric imaging, where the latter could get by with a smaller sample size (assuming reasonable effect sizes that you have boosted through integrating considerations outlined in the sections above). For instance, predictive models could be more useful for early intervention and precision medicine efforts in psychiatry, whereas traditional univariate and multivariate statistical tests could be more appropriate when assessing the interplay between different factors (which could be informed by predictive models) that may explain the variance in treatment outcomes among already diagnosed patients. It should also be noted that predictive models may not always provide any additional relative value in variance explained compared to explanatory models, particularly when effect sizes are very small to begin with.

Taking into consideration context and other external variables contributing to inter-individual variability in brain function may also help us reach towards more generalizable models with clinical utility. For instance, a novel paradigm interleaving task fMRI and transcranial magnetic stimulation (TMS) showcased that TMS effects on subgenual anterior cingulate cortical stimulation were dependent on the brain state elicited in the participant [81]. TMS was more effective in specific brain states engaged through the n-back task, tapping into working memory-related cognition, compared to the weaker effects observed in the absence of a task. Even personality traits can impact the strength of task fMRI-based brain-behavior associations for some cognitive tasks [82], emphasizing the importance of placing participants in an experimental paradigm that intentionally engages targeted cognitive processes, which is not the case in commonly used resting state imaging protocols.

In sum, prioritizing generalizability and improving our experimental study designs and measured outcomes will be essential to successfully attain clinically translational applications of multivariate prediction. Diversification of our samples is also key to this enterprise, given that data collection efforts have been dominated by western, educated, industrialized, rich, and democratic ('WEIRD') samples [83]. It can be particularly challenging to reach this level of generalizability when underrepresented minority groups are less likely to seek and/or access healthcare resources [84–86]. Stepping beyond convenience samples will require much larger changes in infrastructure that need to be prioritized and incentivized by our funding mechanisms and research institutions.

Indeed, we are seeing strides in the right direction with imaging datasets incorporating participants from diverse backgrounds, such as the Healthy Brain Network [87], the ABCD Study [88], and the BrainLat Project [89], and datasets curated by globally-driven initiatives including the Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) Consortium [90], the Global Brain Initiative [91], and the African Brain Disorders Research Network [92]. These examples of consortia-level efforts offer invaluable resources from which smaller studies can draw upon, with specific use cases described in the section below.

### LEVERAGING LARGE SAMPLES TO INFORM SMALLER PSYCHIATRIC SAMPLES

There are certain research questions, particularly within the realm of prediction, where large samples are crucial to avoid some of the pitfalls of multivariate methods that have been alluded to. Although large samples can be hard to come by in psychiatry, there are plenty of existing large datasets that could be leveraged to inform smaller studies [8], depending on the research question you are trying to boost power for. The above-mentioned ABCD Study holds relevance for investigators working in psychiatry, given its sampling of a large neurodevelopmental sample in an age range where many psychiatric disorders tend to onset. The ABCD Study offers a wealth of data relevant for risk factors and onset of psychiatric disorders that researchers around the world working in psychiatry-related fields have been able to leverage in their own research programs.

Our group has recently shown that by applying multivariate methods to various neuroimaging modalities available in the ABCD Study, we can capture lower dimensional patterns of structural and functional brain architecture that correlate robustly with cognitive phenotypes and can be detected with only 41 individuals in the replication sample for working memory-related functional MRI, and  $\sim 100$  subjects for structural MRI [8]. This work also highlights analyses testing out-of-sample prediction performance given different discovery sample sizes, showing that even with 50 subjects in the discovery sample, only  $\sim 98$  subjects are required in replication to predict general cognition from task fMRI data using multivariate methods. Altogether, these results emphasize that thousands of individuals are not required in either discovery or replication samples in pursuit of meaningful and reproducible brain-behavior associations, and shine a more positive light on the sample sizes that many researchers in psychiatry are in reality working with. Others have also commented on the increased predictive utility and reliability of task fMRI for behavioral outcomes when applying multivariate methods [93, 94].

This analytical framework shows promise in being extended to prediction in independent datasets. Recent work has shown that brain-based prediction of cognition in the UK Biobank dataset ( $N \sim 37,000$ ) can be successfully applied to predict cognitive performance in three external transdiagnostic clinical datasets with just hundreds of subjects in each of these samples [78]. There are several other use cases of leveraging large neuroimaging studies that show promise in smaller clinical samples, inspired by examples in other fields. For instance, the genetics field has shown the successful transfer of using discoveries derived from large genomic initiatives (e.g., genome-wide association studies) to smaller psychiatric samples (e.g., through the calculation of polygenic risk scores). The neuroimaging field has incorporated similarly motivated frameworks of the polygenic risk score to neuroimaging-derived summary scores (e.g., polyvertex/voxel scores [63]; polyneuro risk scores [95]).

Genetically informed brain patterns derived in larger samples for known psychiatric genetic risk factors may help guide targets for analysis in smaller samples. The complement component C4 gene is a well-documented risk factor for schizophrenia, with

various levels of evidence pointing to C4's role in synaptic pruning within brain tissue [96–98]. The C4A haplotype has been imputed in large population-based samples, including individuals without known psychiatric or neurological conditions. Findings in tens of thousands of adults from the UK Biobank have shown that C4A is associated with regional cortical anatomy (surface area and thickness) and specific cognitive domains, encompassing reasoning, processing speed, and memory performance [99]. This approach was also extended to youth aged 9–12 years in the ABCD Study, showcasing the impact of C4A on entorhinal cortical morphology in late childhood, which in turn predicted psychotic-like experiences in early adolescence [100]. These regional neuroanatomical effects discovered in large samples could be informative targets for investigations of more deeply phenotyped, but smaller, clinical samples of patients with schizophrenia and related psychotic disorders.

Such genetically informed brain patterns have also shown utility in uncovering novel associations with previously unlinked neuropsychiatric symptom profiles. For example, hereditary hemochromatosis, a disorder characterized by excessive bodily iron levels, has been strongly linked to homozygosity for p.C282Y. Importantly, the clinical course of hereditary hemochromatosis can be remedied by early intervention and treatment [101]. By leveraging UK Biobank data, Loughnan and colleagues [102] demonstrated that carrying p.C282Y, the risk gene for “excessive iron”, is robustly linked to neuroimaging markers of iron deposition in subcortical and cerebellar motor circuits, and disproportionately found in males with diagnosed movement disorders. In a follow-up study, this brain pattern of p.C282Y genetic risk for hereditary hemochromatosis was applied to generate a summary brain score to highlight increased risk for Parkinson's Disease [103]. Motor symptoms are a core feature of many psychiatric diagnoses; thus this work showcases a framework that could be readily extensible to other neuropsychiatric conditions.

Borrowing from medical practices, large-scale initiatives can also be leveraged to derive developmental benchmarks, serving as an anchoring point for smaller samples to be compared against. A landmark study included over 100,000 individuals spanning over 100 studies to create a reference for brain imaging metrics over the lifespan, akin to developmental growth charts [104]. These lifespan brain charts were accompanied by an openly available tool (<http://www.brainchart.io/>) offering researchers the ability to benchmark scans from their own studies. This is of particular interest to researchers and clinicians working in psychiatry, as many psychiatric disorders have been conceptualized as disorders of neurodevelopment, and show notable deviations from normative age trajectories [105, 106].

The aforementioned examples reflect the potential role large datasets could play in shaping the future landscape of psychiatric neuroimaging research. However, findings from consortia-level studies have not yet been ported over to the clinic, as they are often not collecting data that is relevant for real-world outcomes for a psychiatric population of interest. This is reflected by the observed small effect sizes in the prediction of phenotypes related to psychopathology in large datasets, even with multivariate methods [1, 8, 107]. This may be due, in part, to the lower predictive power of neuroimaging measures on self-reported measures, compared to measures obtained more objectively through cognitive tasks [107]. Measures of psychopathology may be more valid in smaller datasets if assessed by a clinician, but this once again poses a challenge for the described framework of generalizing findings from large datasets to clinical samples. A meta-matching approach has shown some success in this domain in the prediction of different, yet correlated, outcome measures across datasets [108]. However, if effect sizes are small to begin with in the training dataset, this will hinder progress toward generalizability. In such cases,

consideration of translating models from smaller samples enriched for brain-behavior associations in a psychiatric sample to larger samples may be worthwhile, which will be discussed in the section below.

### REVERSING THE PREDICTIVE FRAMEWORK: LEARNING FROM SMALLER SAMPLES

‘Big data’ does not necessarily need to come from large sample sizes alone. A wealth of information can be gleaned from smaller but more deeply phenotyped ‘boutique’ samples. There is a growing appreciation for studies that prioritize collection of more data per individual [30, 32]. These precision mapping efforts have been particularly fruitful in the realms of functional MRI [33–36] and have helped define new neuroimaging features that better map onto individual differences. Increased sampling of within-individual imaging data, both within a scan session and across time, may also improve efforts to map neurobiological data to densely sampled symptom and behavioral data, such as through ecological momentary assessments (EMA) [32, 109]. Readers are encouraged to read a recent review [31] covering the strengths and limitations of methodological frameworks incorporating both imaging and EMA data from 2019–2022, and recommendations for psychiatric research.

One example where dense neuroimaging sampling has shown promise in improving prediction is in major depressive disorder. A recent sobering report showcased that in a systematic evaluation of machine learning methods to predict depression status (with a sample of >800 patients), no imaging modality surpassed 62% prediction accuracy in differentiating patients from healthy controls [110]. However, another study painted a much more promising picture by harnessing precision functional mapping data in this patient population [111]. Specifically, Lynch and colleagues (2023) used >21,000 min of longitudinal resting state fMRI data from 187 individuals combined across 11 datasets, showcasing that frontostriatal connectivity patterns were tightly coupled to depressive symptom fluctuations in patients. They found that an expanded frontostriatal network robustly differentiated patients with major depressive disorder from non-clinical controls, and this was replicated in a larger group-level analysis of 1231 subjects. This promising application of using precision functional mapping in a smaller sample size and generalizing to a larger cohort may be particularly useful to validate clinical prediction models of transient states or symptom profiles [32].

Many recently described precision mapping protocols entail hours of data collection on a single individual (some of which are openly available, such as ‘the Midnight Scan Club’ [33]), which may not be an optimal protocol for many patient populations. However, any efforts in increasing the amount of data per individual through longitudinal study designs will likely benefit generalization of prediction models, and ultimately efforts towards precision psychiatry. One group of investigators found that even with standard acquisitions of ~10 min of resting state fMRI data collected at two timepoints (at baseline and 8-weeks after initiating antipsychotic treatment), individualized functional connectivity patterns could better predict individual positive symptoms after 8 weeks of antipsychotic treatment compared to a group-level approach. Similar patterns have been found with longitudinal ‘pre- and post-treatment’ functional connectivity data in major depressive disorder [112]. Task fMRI-based measures also show promise in prediction of symptoms early in the course of treatment; for instance, frontoparietal activation during a cognitive control task significantly predicted improvement in symptoms over a 1-year follow-up period in patients with a recent onset of psychosis [113]. It will be important to increase the number of timepoints and scan times, if possible, in future extensions of these individualized predictive models.

## FUTURE RESEARCH DIRECTIONS AND CONCLUSIONS

It would be a disservice to the field to declare that a sample size of “x” participants is sufficient for brain-behavior associations. The answer lies within many decision points and factors in analysis that we have woven throughout this review, and ultimately boils down to the effect sizes that investigators are measuring. Paying increased attention to choices related to data inclusion and modeling, data acquisition, and analytic frameworks may help boost effect sizes and enhance generalizability, which would pave a more confident path towards clinical translation of neuroimaging-based prediction and enhancing understanding of treatment targets and their underlying mechanisms.

Psychiatric neuroimaging has seen a flurry of discussion around effect sizes and recommended samples. However, simply boosting effect sizes and increasing sample sizes will not be sufficient for truly translational efforts towards precision psychiatry. This will require concerted efforts and incentivization to operationalize and collect more ecologically valid phenotypes. This has been outlined as a key priority for some federally funded initiatives, for instance through the instantiation of the Research Domain Criteria (RDoc) framework by the National Institutes of Mental Health initiative over a decade ago to move from diagnostic labels to neurobiologically-defined constructs [114, 115].

Rather than focusing primarily on increasing sample sizes, we encourage researchers to shift their focus to improving the *quality* of their chosen neuroimaging and clinical measures, obtaining repeated measurements where possible, and leveraging existing datasets to boost effect sizes. Researchers working within psychiatry are encouraged to identify “modifiable” features within their study protocol, which can help increase effect sizes and thus power, without the need to increase sample size. This could be as simple as prioritizing random sampling of both high and low ends of a symptom scale of interest to increase measured variability. Researchers should also prioritize generalizability of their findings to other samples, through the integration of more diverse and inclusive recruitment and retention strategies. We have highlighted several methodological frameworks that may pose additional complexities in their implementation, but researchers specializing in psychiatry need not walk this road alone. Many authors are openly sharing their code and many of the large datasets discussed here are working within an open science framework. Enhancing visibility, accessibility, and educational resources of these readily available tools will enhance the application of methodological advances actively being developed in other areas of neuroimaging to the realm of psychiatry. Finally, brain-behavior mapping, regardless of the study design, will only be successful with ongoing discourse between researchers and clinicians, to ensure that the measures and outcomes incorporated into our studies are in line with what is used to make decisions in clinical practice. The wealth of existing data collection efforts, analytical tools, and interdisciplinary efforts hold promise for boosting power in psychiatric neuroimaging studies, for both the big and the small.

## REFERENCES

- Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*. 2022;603:654–60.
- Liu S, Abdellaoui A, Verweij KJH, van Wingen GA. Replicable brain-phenotype associations require large-scale neuroimaging data. *Nat Hum Behav*. 2023;7:1344–56.
- Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365–76.
- Szucs D, Ioannidis JPA. Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage*. 2020;221:117164.
- Finn ES. Is it time to put rest to rest? *Trends Cogn Sci*. 2021;25:1021–32.
- Rosenberg MD, Finn ES. How to establish robust brain-behavior relationships without thousands of individuals. *Nat Neurosci*. 2022;25:835–7.
- Makowski C, Brown TT, Zhao W, Hagler DJ, Parekh P, Garavan H, et al. Leveraging the Adolescent Brain Cognitive Development Study to improve behavioral prediction from neuroimaging in smaller replication samples. *bioRxiv*. 2023. 1 October 2023. <https://doi.org/10.1093/cercor/bhae223>.
- Elliott ML, Knodt AR, Cooke M, Kim MJ, Melzer TR, Keenan R, et al. General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *Neuroimage*. 2019;189:516–32.
- Greene AS, Gao S, Scheinost D, Constable RT. Task-induced brain state manipulation improves prediction of individual traits. *Nat Commun*. 2018;9:2807.
- Jiang R, Zuo N, Ford JM, Qi S, Zhi D, Zhuo C, et al. Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *Neuroimage*. 2020;207:116370.
- Finn ES, Bandettini PA. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *Neuroimage*. 2021;235:117963.
- Zhao W, Makowski C, Hagler DJ, Garavan HP, Thompson WK, Greene DJ, et al. Task fMRI paradigms may capture more behaviorally relevant information than resting-state functional connectivity. *Neuroimage*. 2023;270:119946.
- Omidvarnia A, Sasse L, Larabi DI, Raimondo F, Hoffstaedter F, Kasper J, et al. Is resting state fMRI better than individual characteristics at predicting cognition? *bioRxiv*. 2023:2023.02.18.529076v4.
- Finn ES, Scheinost D, Finn DM, Shen X, Papademetris X, Constable RT. Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage*. 2017;160:140–51.
- Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K. Toward a ‘treadmill test’ for cognition: Improved prediction of general cognitive ability from the task activated brain. *Hum Brain Mapp*. 2020;41:3186–97.
- Goossen B, van der Starre J, van der Heiden C. A review of neuroimaging studies in generalized anxiety disorder: ‘so where do we stand?’. *J Neural Transm*. 2019;126:1203–16.
- Finn E. To improve big data, we need small-scale human imaging studies. *The Transmitter*. 2024.
- Richtel M. Brain-Imaging Studies Hampered by Small Data Sets, Study Finds. *The New York Times*. 2022.
- Tejavibulya L, Rolison M, Gao S, Liang Q, Peterson H, Dadashkarimi J, et al. Predicting the future of neuroimaging predictive models in mental health. *Mol Psychiatry*. 2022;27:3129–37.
- Tiego J, Martin EA, DeYoung CG, Hagan K, Cooper SE, Pasion R, et al. Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nat Ment Health*. 2023;1:304–15.
- Kang K, Seidlitz J, Bethlehem RAI, Schildcrout J, Tao R, Xiong J, et al. Study design features that improve effect sizes in cross-sectional and longitudinal brain-wide association studies. *bioRxiv*. 2024:2023.05.29.542742v3.
- Haukvik UK, Hartberg CB, Nerland S, Jørgensen KN, Lange EH, Simonsen C, et al. No progressive brain changes during a 1-year follow-up of patients with first-episode psychosis. *Psychol Med*. 2016;46:589–98.
- Roiz-Santiañez R, de la Foz VO-G, Ayesa-Arriola R, Tordesillas-Gutiérrez D, Jorge R, Varela-Gómez N, et al. No progression of the alterations in the cortical thickness of individuals with schizophrenia-spectrum disorder: a three-year longitudinal magnetic resonance imaging study of first-episode patients. *Psychol Med*. 2015;45:2861–71.
- Nesvåg R, Bergmann Ø, Rimol LM, Lange EH, Haukvik UK, Hartberg CB, et al. A 5-year follow-up study of brain cortical and subcortical abnormalities in a schizophrenia cohort. *Schizophr Res*. 2012;142:209–16.
- Makowski C, Bodnar M, Malla AK, Jooper B, Lepage M. Age-related cortical thickness trajectories in first episode psychosis patients presenting with early persistent negative symptoms. *NPJ Schizophr*. 2016;2:16029.
- Makowski C, Bodnar M, Shenker JJ, Malla AK, Jooper B, Chakravarty MM, et al. Linking persistent negative symptoms to amygdala-hippocampus structure in first-episode psychosis. *Transl Psychiatry*. 2017;7:e11195–e1195.
- Raznahan A, Shaw PW, Lerch JP, Clasen LS, Greenstein D, Berman R, et al. Longitudinal four-dimensional mapping of subcortical anatomy in human development. *Proc Natl Acad Sci*. 2014;111:1592–7.
- Walhovd KB, Fjell AM, Giedd J, Dale AM, Brown TT. Through thick and thin: a need to reconcile contradictory results on trajectories in human cortical development. *Cereb Cortex*. 2017;27:1472–81.
- Yip SW, Konova AB. Densely sampled neuroimaging for maximizing clinical insight in psychiatric and addiction disorders. *Neuropsychopharmacology*. 2022;47:395–6.
- McGowan AL, Sayed F, Boyd ZM, Jovanova M, Kang Y, Speer ME, et al. Dense sampling approaches for psychiatry research: combining scanners and smart-phones. *Biol Psychiatry*. 2023;93:681–9.



32. Kraus B, Zinbarg R, Braga RM, Nusslock R, Mittal VA, Gratton C. Insights from personalized models of brain and behavior for identifying biomarkers in psychiatry. *Neurosci Biobehav Rev.* 2023;152:105259.
33. Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, et al. Precision functional mapping of individual human brains. *Neuron.* 2017;95:791–807.e7.
34. Marek S, Greene DJ. Precision functional mapping of the subcortex and cerebellum. *Curr Opin Behav Sci.* 2021;40:12–18.
35. Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, et al. Functional system and areal organization of a highly sampled individual human brain. *Neuron.* 2015;87:657–70.
36. Gratton C, Kraus BT, Greene DJ, Gordon EM, Laumann TO, Nelson SM, et al. Defining individual-specific functional neuroanatomy for precision psychiatry. *Biol Psychiatry.* 2020;88:28–39.
37. Ooi LQR, Orban C, Nichols TE, Zhang S, Tan TWK, Kong R, et al. MRI economics: balancing sample size and scan duration in brain wide association studies. *bioRxiv.* 2024. 18 February 2024. <https://doi.org/10.1101/2024.02.16.580448>.
38. Pardoe HR, Kucharsky Hiess R, Kuzniecky R. Motion and morphometry in clinical and nonclinical populations. *Neuroimage.* 2016;135:177–85.
39. Kong X-Z, Zhen Z, Li X, Lu H-H, Wang R, Liu L, et al. Individual differences in impulsivity predict head motion during magnetic resonance imaging. *PLoS One.* 2014;9:e104989.
40. Ducharme S, Albaugh MD, Nguyen T-V, Hudziak JJ, Mateos-Pérez JM, Labbe A, et al. Trajectories of cortical thickness maturation in normal brain development—the importance of quality control procedures. *Neuroimage.* 2016;125:267–79.
41. Savalia NK, Agres PF, Chan MY, Feczko EJ, Kennedy KM, Wig GS. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Hum Brain Mapp.* 2017;38:472–92.
42. Baum GL, Roalf DR, Cook PA, Ciric R, Rosen AFG, Xia C, et al. The impact of in-scanner head motion on structural connectivity derived from diffusion MRI. *Neuroimage.* 2018;173:275–86.
43. Satterthwaite TD, Wolf DH, Loughhead J, Ruparel K, Elliott MA, Hakonarson H, et al. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage.* 2012;60:623–32.
44. Makowski C, Lepage M, Evans AC. Head motion: the dirty little secret of neuroimaging in psychiatry. *J Psychiatry Neurosci.* 2019;44:62–68.
45. Miezin FM, Maccotta L, Ollinger JM, Petersen SE, Buckner RL. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage.* 2000;11:735–59.
46. Buxton RB, Uludağ K, Dubowitz DJ, Liu TT. Modeling the hemodynamic response to brain activation. *Neuroimage.* 2004;23:S220–S233.
47. Chen G, Taylor PA, Reynolds RC, Leibenluft E, Pine DS, Brotman MA, et al. BOLD Response is more than just magnitude: improving detection sensitivity through capturing hemodynamic profiles. *Neuroimage.* 2023;277:120224.
48. Burock MA, Dale AM. Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient and unbiased approach. *Hum Brain Mapp.* 2000;11:249–60.
49. Dale AM, Buckner RL. Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapp.* 1997;5:329–40.
50. Dale AM. Optimal experimental design for event-related fMRI. *Hum Brain Mapp.* 1999;8:109–14.
51. Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci.* 2017;20:299–303.
52. Marzi C, Giannelli M, Barucci A, Tessa C, Mascalchi M, Diciotti S. Efficacy of MRI data harmonization in the age of machine learning: a multicenter study across 36 datasets. *Sci Data.* 2024;11:115.
53. Pinto MS, Paoletta R, Billiet T, Van Dyck P, Guns P-J, Jeurissen B, et al. Harmonization of brain diffusion MRI: concepts and methods. *Front Neurosci.* 2020;14:396.
54. El-Gazzar A, Thomas RM, van Wingen G. Harmonization techniques for machine learning studies using multi-site functional MRI data. *bioRxiv.* 2023:2023.06.14.544758.
55. Heymans MW, Twisk JWR. Handling missing data in clinical research. *J Clin Epidemiol.* 2022;151:185–8.
56. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009;338:b2393.
57. Allison PD. *The SAGE handbook of quantitative methods in psychology.* United Kingdom: SAGE Publications; 2009. p. 72–89.
58. Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry.* 2005;44:1230–40.
59. Gard AM, Hyde LW, Heeringa SG, West BT, Mitchell C. Why weight? Analytic approaches for large-scale population neuroscience data. *Dev Cogn Neurosci.* 2023;59:101196.
60. Ricard JA, Parker TC, Dhamala E, Kwasa J, Allsop A, Holmes AJ. Confronting racially exclusionary practices in the acquisition and analyses of neuroimaging data. *Nat Neurosci.* 2023;26:4–11.
61. van Buuren S. *Flexible imputation of missing data*, 2nd edition. United Kingdom: CRC Press; 2018.
62. Palmer CE, Zhao W, Loughnan R, Zou J, Fan CC, Thompson WK, et al. Distinct regionalization patterns of cortical morphology are associated with cognitive performance across different domains. *Cereb Cortex.* 2021;31:3856–71.
63. Zhao W, Palmer CE, Thompson WK, Chaarani B, Garavan HP, Casey BJ, et al. Individual differences in cognitive performance are better predicted by global rather than localized BOLD activity patterns across the cortex. *Cereb Cortex.* 2021;31:1478–88.
64. van der Meer D, Frei O, Kaufmann T, Shadrin AA, Devor A, Smeland OB, et al. Understanding the genetic determinants of the brain with MOSTest. *Nat Commun.* 2020;11:3512.
65. Noble S, Curtiss J, Pessoa L, Scheinost D. The tip of the iceberg: a call to embrace anti-localizationism in human neuroscience research. *Imaging Neurosci.* 2024;2:1–10.
66. Abdallah CG, Sheth SA, Storch EA, Goodman WK. Brain imaging in psychiatry: time to move from regions of interest and interpretive analyses to connectomes and predictive modeling? *Am J Psychiatry.* 2023;180:17–19.
67. Spisak T, Bingle U, Wager TD. Multivariate BWAS can be replicable with moderate sample sizes. *Nature.* 2023;615:E4–E7.
68. Botvinik-Nezer R, Wager TD. Reproducibility in neuroimaging analysis: challenges and solutions. *Biol Psychiatry Cogn Neurosci Neuroimaging.* 2023;8:780–8.
69. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun.* 2024;15:1829.
70. Petschnner FH. Practical challenges for precision medicine. *Science.* 2024;383:149–50.
71. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage.* 2017;145:166–79.
72. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage.* 2018;180:68–77.
73. Whelan R, Garavan H. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol Psychiatry.* 2014;75:746–8.
74. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev.* 2015;57:328–49.
75. Thölke P, Mantilla-Ramos Y-J, Abdelhedi H, Maschke C, Dehgan A, Harel Y, et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *Neuroimage.* 2023;277:120253.
76. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci.* 2017;12:1100–22.
77. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorgieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. *Science.* 2024;383:164–7.
78. Chopra S, Dhamala E, Lawhead C, Ricard J, Orcharid E, An L, et al. 252. Reliable and generalizable brain-based predictions of cognitive functioning across common psychiatric illness. *Biol Psychiatry.* 2023;93:5195.
79. Rosenblatt M, Tejavibulya L, Camp CC, Jiang R, Westwater ML, Noble S, et al. Power and reproducibility in the external validation of brain-phenotype predictions. *bioRxiv.* 2023. 30 October 2023. <https://doi.org/10.1101/2023.10.25.563971>.
80. Nielsen AN, Barch DM, Petersen SE, Schlaggar BL, Greene DJ. Machine learning with neuroimaging: evaluating its applications in psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging.* 2020;5:791–8.
81. Grosshagauer S, Wolejt M, Vasileiadi M, Linhardt D, Nohava L, Schuler A-L, et al. Chronometric TMS-fMRI of personalized left dorsolateral prefrontal target reveals state-dependency of subgenual anterior cingulate cortex effects. *Mol Psychiatry.* 2024. 26 March 2024. <https://doi.org/10.1038/s41380-024-02535-3>.
82. Hardikar S, McKeown B, Turnbull A, Xu T, Valk SL, Bernhardt BC, et al. Personality traits vary in their association with brain activity across situations. *bioRxiv.* 2024:2024.04.18.590056.
83. Jones D. Psychology. A WEIRD view of human nature skews psychologists' studies. *Science.* 2010;328:1627.
84. Choi SW, Ramos C, Kim K, Azim SF. The association of racial and ethnic social networks with mental health service utilization across minority groups in the USA. *J Racial Ethn Health Disparities.* 2019;6:836–50.
85. Lu W, Todhunter-Reid A, Mitsdarffer ML, Muñoz-Laboy M, Yoon AS, Xu L. Barriers and facilitators for mental health service use among racial/ethnic minority adolescents: a systematic review of literature. *Front Public Health.* 2021;9:641605.

86. Kim SB, Lee YJ. Factors associated with mental health help-seeking among Asian Americans: a systematic review. *J Racial Ethn Health Disparities*. 2022;9:1276–97.
87. Alexander LM, Escalera J, Ai L, Andreotti C, Febre K, Mangone A, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data*. 2017;4:170181.
88. Volkow ND, Koob GF, Croyle RT, Bianchi DW, Gordon JA, Koroshetz WJ, et al. The conception of the ABCD study: from substance use to a broad NIH collaboration. *Dev Cogn Neurosci*. 2018;32:4–7.
89. Prado P, Medel V, Gonzalez-Gomez R, Sainz-Ballesteros A, Vidal V, Santamaría-García H, et al. The BrainLat project, a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds. *Sci Data*. 2024;11:19.
90. Thompson PM, Jahanshad N, Ching CRK, Salminen LE, Thomopoulos SI, Bright J, et al. ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry*. 2020;10:100.
91. Valdes-Sosa PA, Galan-García L, Bosch-Bayard J, Bringas-Vega ML, Aubert-Vazquez E, Rodriguez-Gil I, et al. The Cuban Human Brain Mapping Project, a young and middle age population-based EEG, MRI, and cognition dataset. *Sci Data*. 2021;8:45.
92. Aborode AT, Idowu NJ, Tundealao S, Jaiyeola J, Ogunware AE. Strengthening brain research in Africa. *J Alzheimers Dis Rep*. 2023;7:989–92.
93. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, et al. What is the test-retest reliability of common task-functional MRI measures? new empirical evidence and a meta-analysis. *Psychol Sci*. 2020;31:792–806.
94. Dubois J, Adolphs R. Building a science of individual differences from fMRI. *Trends Cogn Sci*. 2016;20:425–43.
95. Byington N, Grimsrud G, Mooney MA, Cordova M, Doyle O, Hermsillo RJM, et al. Polyneuro risk scores capture widely distributed connectivity patterns of cognition. *Dev Cogn Neurosci*. 2023;60:101231.
96. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016;530:177–83.
97. Sellgren CM, Gracias J, Watmuff B, Biag JD, Thanos JM, Whittredge PB, et al. Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nat Neurosci*. 2019;22:374–85.
98. Sellgren CM, Sheridan SD, Gracias J, Xuan D, Fu T, Perlis RH. Patient-specific models of microglia-mediated engulfment of synapses and neural progenitors. *Mol Psychiatry*. 2017;22:170–7.
99. O'Connell KS, Sønderby IE, Frei O, van der Meer D, Athanasiu L, Smeland OB, et al. Association between complement component 4A expression, cognitive performance and brain imaging measures in UK Biobank. *Psychol Med*. 2021;52:1–11.
100. Hernandez LM, Kim M, Zhang P, Bethlehem RAI, Hoftman G, Loughnan R, et al. Multi-ancestry phenome-wide association of complement component 4 variation with psychiatric and brain phenotypes in youth. *Genome Biol*. 2023;24:42.
101. Powell LW, Seckington RC, Deugnier Y. Haemochromatosis. *Lancet*. 2016;388:706–16.
102. Loughnan R, Ahern J, Tompkins C, Palmer CE, Iversen J, Thompson WK, et al. Association of genetic variant linked to hemochromatosis with brain magnetic resonance imaging measures of iron and movement disorders. *JAMA Neurol*. 2022;79:919–28.
103. Loughnan R, Ahern J, Boyle M, Jernigan TL, Donald J Hagler J, et al. Neural archetypes learnt from hemochromatosis reveals iron dysregulation in motor circuits. *medRxiv*. 2024. 2022.10.22.22281386v3. <https://doi.org/10.1101/2022.10.22.22281386>.
104. Bethlehem RAI, Seidlitz J, White SR, Vogel JW, Anderson KM, Adamson C, et al. Brain charts for the human lifespan. *Nature*. 2022;604:525–33.
105. Marín O. Developmental timing and critical windows for the treatment of psychiatric disorders. *Nat Med*. 2016;22:1229–38.
106. Uhlhaas PJ, Davey CG, Mehta UM, Shah J, Torous J, Allen NB, et al. Towards a youth mental health paradigm: a perspective and roadmap. *Mol Psychiatry*. 2023. 14 August 2023. <https://doi.org/10.1038/s41380-023-02202-z>.
107. Kong R, Yang Q, Gordon E, Xue A, Yan X, Orban C, et al. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb Cortex*. 2021;31:4477–4500.
108. He T, An L, Chen P, Chen J, Feng J, Bzdok D, et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat Neurosci*. 2022;25:795–804.
109. Kraus B, Sampathgiri K, Mittal VA. Accurate machine learning prediction in psychiatry needs the right kind of information. *JAMA Psychiatry*. 2024;81:11–12.
110. Winter NR, Blanke J, Leenings R, Ernsting J, Fisch L, Sarink K, et al. A systematic evaluation of machine learning-based biomarkers for major depressive disorder. *JAMA Psychiatry*. 2024;81:386–95.
111. Lynch CJ, Elbau I, Ng T, Ayaz A, Zhu S, Manfredi N, et al. Expansion of a frontostriatal salience network in individuals with depression. *bioRxiv*. 2023. 14 August 2023. <https://doi.org/10.1101/2023.08.09.551651>.
112. Zhao Y, Dahmani L, Li M, Hu Y, Ren J, Lui S, et al. Individualized functional connectome identified replicable biomarkers for dysphoric symptoms in first-episode medication-naïve patients with major depressive disorder. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2023;8:42–51.
113. Smucny J, Lesh TA, Carter CS. Baseline frontoparietal task-related BOLD activity as a predictor of improvement in clinical symptoms at 1-year follow-up in recent-onset psychosis. *Am J Psychiatry*. 2019;176:839–45.
114. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med*. 2013;11:126.
115. Cuthbert BN. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*. 2014;13:28–35.

## ACKNOWLEDGEMENTS

The authors would like to thank Timothy Brown, Terry Jernigan, and Robert Loughnan for insightful conversations that helped shape some of the perspectives shared in this manuscript.

## AUTHOR CONTRIBUTIONS

CM and AMD conceived the ideas presented in this manuscript. CM wrote the first draft. TN provided statistical expertise on presented concepts. All authors contributed to editing the final manuscript.

## FUNDING

This work was supported by the National Institutes of Mental Health (award number K99MH132886; CM). The funding agency did not influence the perspectives outlined in the manuscript, nor the decision to publish.

## COMPETING INTERESTS

AMD reports that he was a Founder of and holds equity in CorTechs Labs, Inc., and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc. He receives funding through research grants from GE Healthcare to UCSD. The terms of these arrangements have been reviewed by and approved by UCSD in accordance with its conflict of interest policies. AMD also reports that he has memberships with the following research consortia: Alzheimer's Disease Genetics Consortium (ADGC); Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA); Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL); Psychiatric Genomics Consortium (PGC). All other authors have no conflicts of interest.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Carolina Makowski.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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