How to establish robust brain–behavior relationships without thousands of individuals

Can studying individual differences in brain structure and function reveal individual differences in behavior? Analyses of MRI data from nearly 50,000 individuals may suggest that the possibility is fleeting. Although sample size is important for brain-based prediction, researchers can take other steps to build better biomarkers. These include testing model generalizability across people, datasets, and time points and maximizing model robustness by optimizing brain data acquisition, behavioral measures, and prediction approaches.

Monica D. Rosenberg and Emily S. Finn

In a recent *Nature* paper, Marek, Tervo-Clemmens, and colleagues[1] investigated whether individual differences in brains predict individual differences in behavior. To do so, they analyzed publicly available MRI and behavioral data from thousands of volunteers, correlating tens of thousands of measures of brain function (functional connections measured with functional MRI) and structure (cortical thickness estimates measured with structural MRI) with dozens of measures of cognition and psychopathology. Although they identified relationships between brain measures and behavior, these correlations were unlikely to replicate unless they were defined using MRI data from thousands of individuals. Furthermore, the strongest correlations — which are presumably the most likely to be published — were the least likely to replicate in new data.

This paper has captured attention not because its results are controversial or surprising. On the contrary, they are rigorous and guaranteed any time mass univariate tests are used to correlate many measures with weak ground-truth relationships. Rather, the results generated discussion in the field and popular press partly because, although the authors generally took care to limit the scope of their work to what they coined 'brain-wide association studies' (BWAS), they have been misinterpreted as undermining MRI research as a whole. Marek et al. emphasize that this isn't true. In this Comment, we elaborate on why it isn't true and explain steps that MRI researchers have taken and can take to identify real brain–behavior associations.

For nearly three decades, MRI research has successfully characterized how brain activity changes as people see, think, and do different things, questions that the BWAS approach does not address. Groundbreaking functional MRI work, for example, compared brain activity evoked by pictures of faces with that evoked by pictures of objects within subjects to identify a face-sensitive patch of cortex in 12 of 15 participants[2] that has since been replicated in nearly every individual tested. Other influential work discovered willful modulation of brain activity by one patient in what otherwise appeared to be a persistent vegetative state[3] and revealed experience-dependent maintenance of functional brain architecture in three volunteers in arm casts[4]. The results of these three studies changed the way we think about the brain using MRI data from only 19 people. Although BWAS analyses as Marek et al. define them require large amounts of data, some of the most well-replicated findings in human neuroscience come from studies that used carefully designed task paradigms to measure well-characterized cognitive processes in a small number of individuals.

This is not to say that sample size is not important for discovering replicable brain–behavior relationships: it is. Underpowered analyses can waste time and money and sow distrust. But more data alone doesn't guarantee better science. The kind of data we collect and the analyses we apply to them also matter. As scientists who care deeply about discovering real insights about the brain and mind, here we spotlight two other paths toward replicable brain–behavior relationships that do not necessarily require thousands of individuals per study. We first emphasize the importance of testing the generalizability of models that predict behavioral variables from brain features. We next suggest ways to build more robust brain-based predictive models in the first place by making thoughtful choices about brain data acquisition, behavioral targets of prediction, and approaches to model building.

Testing model generalizability

We cannot know from sample size alone whether a model that predicts behavior from brain features will generalize; to tell for sure, we must test it on new data. One way to do this is to test for generalizability within a dataset using internal validation approaches such as *k*-fold cross-validation[5,6]. Although internal validation is a fine start, external validation — testing the success of a predictive model in an entirely separate dataset, ideally from an independent data collection site — is a necessary test of population-level model generalizability[7]. Validation sets can be large open-access datasets, such as the Adolescent Brain Cognitive Development Study and Human Connectome Project samples, or smaller datasets shared by individual labs. Differences in imaging parameters and behavioral measures between samples provide opportunities for strong tests of model generalizability. A robust model of a cognitive ability, for example, should generalize to samples with different MRI acquisition parameters and behavioral measures of the same underlying construct. At the same time, model failures can be informative[8]. Age-prediction error, for example, has been related to individual differences in risk preference in young adults[9], and unsuccessful model generalizability across populations could point to differences in brain–behavior relationships when alternative explanations are ruled out. Testing model validity in multiple independent datasets is necessary for developing robust models with real-world relevance.

External model validation can be combined with other approaches to testing the robustness of brain–behavior associations. Preregistration, for example, limits researcher degrees of freedom by 'locking in' hypotheses and analysis methods ahead of time and is gaining popularity

Box 1 | Recommendations for establishing generalizable brain–behavior associations

Test and validate existing models:

- Do out-of-sample prediction rather than within-sample correlation
- Validate models in a single dataset using cross-validation (fine) or across entirely separate datasets (much better)
- Share data and models themselves, so that models can be tested and potentially tweaked on independent data by multiple laboratories
- Preregister hypotheses and analysis methods
- Assess whether models defined to predict individual differences in behavior capture within-subject change
- Test whether experimentally manipulating behavior (for example, with training or pharmacological intervention) results in expected brain changes
- Test whether experimentally manipulating brain signatures (for example, with real-time neurofeedback) results in expected behavioral change

Set models up to generalize:

- Choose behavioral measures with demonstrated reliability and sensitivity to individual differences
- For functional acquisitions, consider task states rather than rest, especially task(s) tailored to the behavior or phenotype of interest
- Use multivariate rather than univariate approaches to model building
- Allow for innovative new behavioral measures and scan paradigms to emerge from smaller-scale studies for eventual inclusion in large-scale consortium efforts

in neuroimaging research[10–12]. Likewise, sharing a published model's features and weights essentially 'preregisters' it by making it available to other groups to test on new data[8]. Real-time neurofeedback offers another way to preregister and test (causal) hypotheses about brain–behavior relationships, as experimenters must decide what brain feature(s) will govern the feedback before data collection[13]. Finally, even without a formal predictive model to assess, researchers can try to replicate previous findings. Incentives such as journal policies that welcome replication and the Organization for Human Brain Mapping's Replication Award are contributing to a growing emphasis on replication. Together with model validation, preregistration, model-sharing, and replication can help to identify robust relationships between individual differences in brain features and behavior.

Although Marek et al. focus on across-subject analyses, researchers can take advantage of growing repositories of deep imaging data[14] to complement individual differences studies with within-subject investigation. Imagine, for example, that we identify a functional brain network whose strength scales with the ability to regulate emotion across individuals. To better understand this relationship, we can ask whether the same network fluctuates with emotion regulation within an individual. Does it vary with states that affect emotion regulation or change with this ability across the lifespan? We can also ask whether manipulating this hypothetical network alters emotion regulation and vice versa. Does training someone to strengthen their network with real-time neurofeedback improve emotion regulation? Does a behavioral intervention that benefits emotion regulation also strengthen the network? Following this logic, recent work identified a functional connectivity-based model that predicts different measures of sustained attention in independent datasets[15–17], fluctuates with attention task performance across minutes, days, and months[18], and is sensitive to pharmacological manipulations that affect attentional state[10,18,19]. Combining across- and within-individual analyses is a powerful way to assess the validity and practical relevance of brain–behavior relationships.

Building generalizable models

To complement efforts to test and validate existing models, we should build new models with an eye toward maximizing generalizability with thoughtful choices during experimental design, data collection, and analysis.

First, we should choose the right behavioral measure(s). Marek et al. test 41 demographic, cognitive, and mental health variables, and focus on cognitive ability and psychopathology. They find more success and generalizability for models that predict cognitive ability, in line with mounting evidence that performance-based variables, rather than self-report questionnaires, are often more robustly predictable from brain data[20,21]. However, this does not necessarily indict self-report as a whole, as the distinction could lie in the constructs being measured and/or their amenability to introspection. Using ecological momentary assessments to densely sample individuals' moods, symptoms, and other experiences could improve the reliability and validity of self-report data[22]. Also of note, tried-and-true performance measures from traditional tasks often suffer from limited variation in the general population, making them unsuitable for research into individual differences[23]. Thus, choosing appropriate behavioral measures is a challenging chicken-and-egg problem: to determine whether individual differences in brain data are meaningful, we must relate them to out-of-scanner behaviors and phenotypes, but without knowing which behaviors and phenotypes are most robustly reflected in biology, we cannot choose the most valid targets. Future research should invest in developing new assays that evince meaningful and stable individual variability in both normative and clinical populations. Targeting brain-based predictions toward longitudinal real-world outcomes is also an important gold standard[8,24].

Second, we should choose the right brain acquisition state. Marek et al. focus largely on resting state. This makes sense, given that large-scale human neuroimaging datasets are currently dominated by resting-state acquisitions, sometimes accompanied by a handful of traditional cognitive tasks. However, an overreliance on resting state may be harming the sensitivity and generalizability of BWASs. To see why, consider analogies from other fields of medicine: to screen for abnormal heart rhythms, rather than measuring heart rate while patients sit on a couch, cardiologists subject them to a treadmill test. To screen for diabetes risk, rather than measure blood sugar when patients come in off the street, physicians conduct a glucose-tolerance test under controlled lab conditions. Similarly, to assess brain function, rather than using unconstrained rest, we should put people under the conditions in which the relevant characteristics and/or vulnerabilities are most likely to emerge. Indeed, task functional MRI data yield more accurate predictions of phenotypes (see ref. [25] for a review), including traditional tasks (for example, *n*-back, emotional faces[26]) as well as naturalistic tasks such as watching a movie[20]. Functional connectivity during task states is often more reliable in and of itself[27,28], which may explain some of its increased sensitivity to behavior. (However, we caution researchers against blindly optimizing for test–retest reliability of brain data alone: not only does increased reliability not necessarily guarantee improved behavior prediction, but also some degree of within-subject change in

brain function is expected and often reflects meaningful processes[29].) Therefore, rather than defaulting to rest, we can improve our models by using thoughtfully chosen task paradigms, perhaps even tailoring them to our phenotype(s) or behavior(s) of interest.

Third, we should focus on patterns of brain features rather than features in isolation. Marek et al. show improved reliability of multivariate patterns over univariate associations. This is perhaps unsurprising given that brain functions are inherently complex and interdependent, and our measures of them are inherently noisy; therefore, because ground-truth effect sizes for individual features are small, we get better signal when aggregating across many features. Another benefit to multivariate approaches is that they eschew the need to correct for multiple comparisons across features, which, as Marek et al. point out, can harm replication and generalizability by increasing false negatives. A drawback of multivariate approaches is that models can be harder to interpret. Furthermore, the models themselves — that is, the weights on specific features — can also be somewhat unreliable, tempering potential interpretations[1,30]. One solution is to use 'virtual lesion' approaches, which selectively remove features (for example, connections from a single canonical brain network) to determine the extent to which model performance suffers. This approach can point to feature sets with above-average importance for predicting the phenotype of interest. Regardless, both theory and empirical evidence suggest that multivariate approaches are much more appropriate than mass univariate tests for establishing generalizable brain–behavior associations. Moving forward, our field can strive to develop the BWAS equivalent of GWAS polygenic risk scores that combine multiple features, or even outputs from multiple models, to produce a combined estimated phenotype prediction.

Conclusions

While large sample sizes certainly help, they are not the only route to generalizable brain–behavior relationships. Here, we have highlighted steps researchers can take to test for and promote replicability even with sample sizes in the tens or perhaps hundreds of individuals (see Box 1).

One danger in discounting the value of smaller-scale datasets based on Marek et al.'s findings is that doing so could stifle innovation. Sample sizes in the thousands or tens of thousands are virtually only achievable in large-scale consortia with major financial and logistical support. While these datasets are highly valuable, they are often limited to the tried-and-true measurements for both behavioral and brain data ('science by committee'), which paradoxically may be some of the worst for studying individual differences[23]. It is risky to include newer and relatively untested behavioral measurements and/or scan paradigms in a protocol destined to be run on thousands of people, but if we do not allow evidence to build up from smaller studies, we risk remaining stuck in local optima for how we acquire data, which will severely hamper our ability to make incisive discoveries about brain–behavior relationships in the long run.

Overall, we reaffirm Marek, Tervo-Clemmens, and colleagues' assertion that small-sample neuroimaging studies have an important place, and furthermore contend that, with careful model building and proper validation, they will continue to have value for discovering robust links between brain and behavior. We can do better without necessarily going bigger. ❐

Monica D. Rosenberg[1,2]✉ and Emily S. Finn[3]✉

[1]*Department of Psychology, The University of Chicago, Chicago, IL, USA.* [2]*Neuroscience Institute, The University of Chicago, Chicago, IL, USA.* [3]*Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA.*
✉*e-mail: mdrosenberg@uchicago.edu; emily.s.finn@dartmouth.edu*

Published online: 16 June 2022
https://doi.org/10.1038/s41593-022-01110-9

References

1. Marek, S. et al. *Nature* **603**, 654–660 (2022).
2. Kanwisher, N., McDermott, J. & Chun, M. M. *J. Neurosci.* **17**, 4302–4311 (1997).
3. Owen, A. M. et al. *Science* **313**, 1402 (2006).
4. Newbold, D. J. et al. *Neuron* **107**, 580–589.e6 (2020).
5. Poldrack, R. A., Huckins, G. & Varoquaux, G. *JAMA Psychiatry* **77**, 534–540 (2020).
6. Scheinost, D. et al. *Neuroimage* **193**, 35–45 (2019).
7. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. *Nat. Neurosci.* **20**, 365–377 (2017).
8. Rosenberg, M. D., Casey, B. J. & Holmes, A. J. *Nat. Commun.* **9**, 589 (2018).
9. Rudolph, M. D. et al. *Dev. Cogn. Neurosci.* **24**, 93–106 (2017).
10. Chamberlain, T. A. & Rosenberg, M. D. *Cereb. Cortex* **2022**, bhac020 (2022).
11. Whitfield-Gabrieli, S. et al. *JAMA Psychiatry* **77**, 378–386 (2020).
12. Ellwood-Lowe, M. E., Whitfield-Gabrieli, S. & Bunge, S. A. *Nat. Commun.* **12**, 7183 (2021).
13. deBettencourt, M. T. & Norman, K. A. *Curr. Biol.* **26**, R673–R675 (2016).
14. Naselaris, T., Allen, E. & Kay, K. *Curr. Opin. Behav. Sci.* **40**, 45–51 (2021).
15. Rosenberg, M. D. et al. *Nat. Neurosci.* **19**, 165–171 (2016).
16. Kardan, O. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.01.454530 (2022)
17. Wu, E. X. W. et al. *Neuroimage* **209**, 116535 (2020).
18. Rosenberg, M. D. et al. *Proc. Natl Acad. Sci. USA* **117**, 3797–3807 (2020).
19. Rosenberg, M. D. et al. *J. Neurosci.* **36**, 9547–9557 (2016).
20. Finn, E. S. & Bandettini, P. A. *Neuroimage* **235**, 117963 (2021).
21. Li, J. et al. *Neuroimage* **196**, 126–141 (2019).
22. Ebner-Priemer, U. W. & Trull, T. J. *Psychol. Assess.* **21**, 463–475 (2009).
23. Hedge, C., Powell, G. & Sumner, P. *Behav. Res. Methods* **50**, 1166–1186 (2018).
24. Yip, S. W., Kiluk, B. & Scheinost, D. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 748–758 (2020).
25. Finn, E. S. *Trends Cogn. Sci.* **25**, 1021–1032 (2021).
26. Greene, A. S., Gao, S., Scheinost, D. & Constable, R. T. *Nat. Commun.* **9**, 2807 (2018).
27. Finn, E. S. et al. *Neuroimage* **160**, 140–151 (2017).
28. Vanderwal, T. et al. *Neuroimage* **157**, 521–530 (2017).
29. Finn, E. S. & Rosenberg, M. D. *Neuroimage* **239**, 118254 (2021).
30. Tian, Y. & Zalesky, A. *Neuroimage* **245**, 118648 (2021).

Competing interests

The authors declare no competing interests.