Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05745-x

Received: 10 April 2022

Accepted: 19 January 2023

Published online: 8 March 2023

Open access

Check for updates

Tamas Spisak[1,2]✉, Ulrike Bingel[2] & Tor D. Wager[3]

ARISING FROM: S. Marek et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* https://doi.org/10.1038/s41586-022-04492-9 (2022)

Brain-wide association studies (BWAS)—which correlate individual differences in phenotypic traits with measures of brain structure and function—have become a dominant method for linking mind and brain over the past 30 years. Univariate BWAS typically test tens to hundreds of thousands of brain voxels individually, whereas multivariate BWAS integrate signals across brain regions into a predictive model. Numerous problems have been raised with univariate BWAS, including a lack of power and reliability and an inability to account for pattern-level information embedded in distributed neural circuits[1–4]. Multivariate predictive models address many of these concerns, and offer substantial promise for delivering brain-based measures of behavioural and clinical states and traits[2,3].

In their recent paper[4], Marek et al. evaluated the effects of sample size on univariate and multivariate BWAS in three large-scale neuroimaging datasets and came to the general conclusion that "BWAS reproducibility requires samples with thousands of individuals". We applaud their comprehensive analysis, and we agree that (1) large samples are needed when conducting univariate BWAS and (2) multivariate BWAS reveal substantially larger effects and are therefore more highly powered.

Marek et al.[4] find that multivariate BWAS provide inflated in-sample associations that often cannot be replicated (that is, are underpowered) unless thousands of participants are included. This implies that effect-size estimates from the discovery sample are necessarily inflated. However, we distinguish between the effect-size estimation method (in-sample versus cross-validated) and the sample (discovery versus replication), and show that, with appropriate cross-validation, the in-sample inflation that Marek et al.[4] report in the discovery sample can be entirely eliminated. With additional analyses, we demonstrate that multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases. Specifically, applying a standard multivariate prediction algorithm to functional connectivity in the Human Connectome Project yielded replicable effects with sample sizes of 75–500 for 5 of 6 phenotypes tested (Fig. 1).

These analyses are limited to a selected number of phenotypes in a relatively high-quality dataset (measured in a young adult population with a single scanner) and should not be overgeneralized. However, they highlight that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship and that, with proper internal validation, appropriate effect-size estimates and sufficiently large effects for moderately sized studies are possible.

Marek et al.[4] evaluate in-sample effect-size inflation in multivariate BWAS by training various multivariate models in a 'discovery sample' and comparing the in-sample effect sizes (prediction–outcome correlation, r) estimated from the training sample to the performance in an independent replication sample. On the basis of a bootstrap analysis, with variously sized pairs of samples drawn randomly from the Adolescent Brain Cognitive Development study, the authors report a severe effect-size inflation of $\Delta r = -0.29$ (average difference between the in-sample effect sizes in the discovery sample and the out-of-sample effect sizes in the replication sample) and conclude that "[e]ven at the largest sample sizes ($n \approx 2{,}000$), multivariate in-sample associations remained inflated on average".

The issue with claims of inflation is that the in-sample effect size estimates of Marek et al.[4] were based on training multivariate models on the entire discovery sample, without cross-validation or other internal validation (as confirmed by inspection of the code and discussion with the authors). Such in-sample correlations are not valid effect-size estimates, as they produce a well-known overfitting bias that increases with model complexity[5]. Standard practice in machine learning is to evaluate model accuracy (and other performance metrics) on data independent of those used for training. In line with current recommendations for multivariate brain–behaviour analyses[6,7], this is typically performed using internal cross-validation (for example, k-fold) to estimate unbiased effect sizes in a discovery sample, and (less commonly) further validating significant cross-validated effects in held-out or subsequently acquired replication samples[2,5].

Using cross-validation to estimate discovery-sample effects impacts the pool of studies selected for replication attempts, the degree of effect-size attenuation in replication samples, and the sample size needed for effective replication and mitigation of publication bias. To demonstrate this and provide quantitative estimates of sample size requirements in multivariate BWAS, we analysed functional connectivity data from the Human Connectome Project[8] (one of the datasets in Marek et al.[4]) using cross-validation to estimate discovery-sample effect sizes. As shown in Fig. 1a–d, cross-validated discovery effect-size estimates are unbiased (that is, not inflated on average), irrespective of the sample size and the magnitude of the effect. As expected, even with cross-validation, smaller sample sizes resulted in lower power (Fig. 1e) and increased variability in effect-size estimates across samples (Fig. 1c). Such variability is undesirable because it reduces the probability of independent replication (Fig. 1f). Moreover, selection biases—most notably, publication bias—can capitalize on such variability to inflate effect sizes in the literature (Fig. 1g). Although these effects of using small sample sizes are undesirable, they do not invalidate

[1]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Medicine Essen, Essen, Germany. [2]Center for Translational Neuro- and Behavioral Sciences, Department of Neurology, University Medicine Essen, Essen, Germany. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. ✉e-mail: tamas.spisak@uk-essen.de

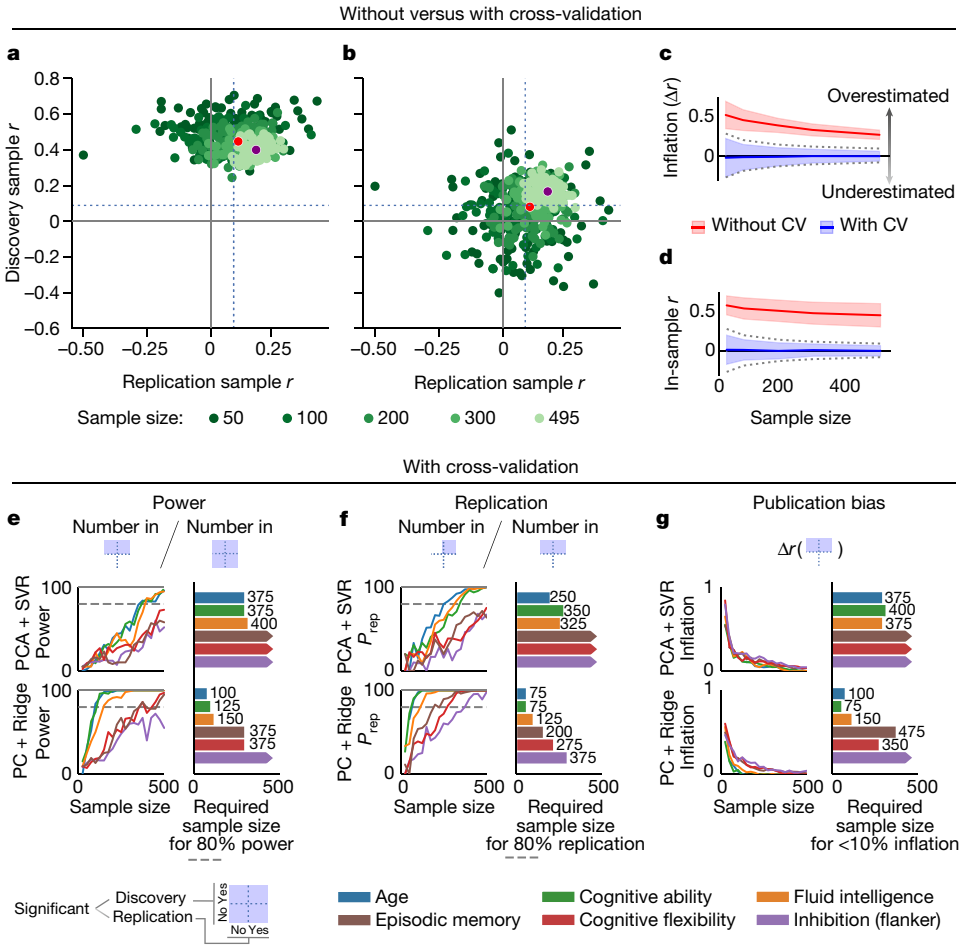

Fig. 1 | Examples of multivariate BWAS providing unbiased effect sizes and high replicability with low to moderate sample sizes. a, Discovery sample effects in multivariate BWAS are inflated only if estimates are obtained without cross-validation (CV). **b**, Cross-validation fully eliminates in-sample effect-size inflation and, as a consequence, provides higher replicability. Data are from the Human Connectome Project (HCP1200, PTN release, $n = 1{,}003$). Each point in **a** and **b** corresponds to one bootstrap subsample, as in figure 4b of Marek et al.[4]. The dotted lines denote the threshold for $P = 0.05$ with $n = 495$. Mean multivariate brain–behavioural phenotype associations across 100 bootstrap samples at $n = 200$ and for the full sample are denoted by red and purple dots. **c**, The inflation of in-sample effect size obtained without cross-validation (red) is reduced, but does not disappear, at higher sample sizes. Conversely, cross-validated estimates (blue) are slightly pessimistic with low sample sizes and become quickly unbiased as sample size is increased. **d**, Without cross-validation, in-sample effect-size estimates are non-zero ($r \approx 0.5$, red), even when predicting permuted outcome data. Cross-validation eliminates systematic bias across all sample sizes (blue). The dashed lines in **c** and **d** denote 95% parametric confidence intervals, and the shaded areas denote bootstrap- and permutation-based confidence intervals. **e**,**f**, Cross-validated analysis reveals that sufficient in-sample power (**e**) and out-of-sample replication probability (P_{rep}) (**f**) can be achieved for a variety of phenotypes at low or moderate sample sizes. 80% power and P_{rep} are achievable in <500 participants for 3 out of 6 phenotypes (coloured bars) using the prediction algorithm of Marek et al.[4] (**e** and **f** (top), the sample size required for 80% power or P_{rep} is shown). The remaining three phenotypes require sample sizes of >500 (bars with arrows). Power and P_{rep} can be substantially improved with a ridge regression-based model recommended in some comparison studies[10,11] (**e** and **f** (bottom), with 80% power and P_{rep} with sample sizes as low as $n = 100$ and $n = 75$, respectively, when predicting cognitive ability, and sample sizes between 75 and 375 for other investigated variables (fluid intelligence, episodic memory and cognitive flexibility), except inhibition assessed with the flanker task, which replicated with $n = 375$ but did not reach 80% power with $n = 500$. **g**, We estimated interactions between sample size and publication bias by computing effect size inflation ($r_{discovery} - r_{replication}$) only for those bootstrap cases in which prediction performance was significant ($P > 0.05$) in the replication sample. Our analysis shows that the effect-size inflation due to publication bias is modest (<10%) with fewer than 500 participants for half of the phenotypes using the model from Marek et al.[4] and all phenotypes but the flanker using the ridge model. The blue squares show conditional relationships assessed to derive metrics in **e**,**f** and **g** with reference to **b**. The top and bottom squares indicate positive and negative results in the discovery sample, respectively. The left and right squares indicate negative and positive results in the replication sample. The blue squares indicate how these conditions were applied to derive the metrics.

the use of multivariate BWAS in small samples, and publication biases can be mitigated by practices that, like internal cross-validation, are quickly becoming standards in the field[2,5]. These include preregistration, registered reports, reporting confidence intervals and the use of hold-out samples tested only once on a single, optimized model to avoid overfitting.

Given these considerations, we wondered how many participants are generally required for multivariate BWAS. The answer to this question depends on the reliability of both phenotypic and brain measures, the size of the effects linking them, the algorithm and model-selection steps used and the use cases for the resulting brain measures. For example, multivariate models trained on as few as 20 participants[9] can have high reliability (ICC = 0.84)[10], broad external validity and large effect sizes (Hedges $g = 2.3$)[11] in independent samples (for example, more than 600 participants from 20 independent studies in ref. [11]) when predicting behavioural states within-person rather than traits. In this case, the benefit of large samples is primarily in accurately estimating local brain weights (model parameters)[12] rather than increasing out-of-sample

accuracy. Here we performed functional connectivity-based multivariate BWAS with cognitive ability (the phenotype shown in figure 4 of Marek et al.[4]) and five other cognition-related example phenotypes selected at random and demonstrate that, even when predicting trait-level phenotypes, as Marek et al.[4] did, sample sizes of 75–500 are sufficient in five out of six of cases that we tested (or three out of six cases using the prediction algorithm of Marek et al.[4]) to achieve high statistical power and replicability (for example, 80%) and to mitigate effect size inflation due to publication bias.

The basis for these estimates is shown in Fig. 1e–g. Using cross-validated discovery sample effect-size estimates, the multivariate BWAS model of Marek et al.[4]—principal-component-based reduction of bivariate connectivity followed by support vector regression (PCA + SVR)—showed 80% in-sample power and 80% out-of-sample replication probability (P_{rep}) at $n < 500$ for three out of six phenotypes that we examined (age, cognitive ability and fluid intelligence). However, this model has been shown to be disadvantageous in some comparison studies[12,13]. We therefore performed the same power and sample-size calculations for a multivariate BWAS using another approach—ridge regression on partial correlation matrices with a default shrinkage parameter of 1 (PC + ridge; Supplementary Methods). Although this approach is still probably sub-optimal[12,13] (we avoided testing other models to avoid overfitting), it substantially improved the power (Fig. 1e (bottom)), independent replication probability ([P_{rep}]; Fig. 1f (bottom)) and resistance to inflation due to publication bias (Fig. 1g (bottom)). Eighty per cent power and P_{rep} were achieved at sample sizes from 75 to 150 for age (included as a reference variable), cognitive ability and fluid intelligence, and sample sizes < 400 for all phenotypes except for inhibition measured by the flanker task (a measure that is known to have low reliability[14]).

Our results highlight, that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship, which subsumes the amount, quality, homogeneity and reliability of both brain and phenotypic measures, and the degree to which a particular brain measure is relevant to a particular phenotype. Effect sizes will probably vary widely across studies; for example, cortical thickness can also reliably predict 4 out of the 6 investigated phenotypes with $n < 500$, although with smaller effect sizes on average (functional connectivity, mean $r = 0.2$; cortical thickness, mean $r = 0.1$; Supplementary Fig. 2). Although our results were derived from a relatively high-quality dataset and used an algorithm expected to yield larger effect sizes than that of Marek et al.[4], they are in agreement with analytical calculations showing that BWAS that explain more than 1% of the phenotype's variance can be replicable with sample sizes below 1,000 (Supplementary Methods). For example, a model that explains $r^2 = 0.01$ (1% of variance) achieves 80% power in a prospective replication with $n = 801$, and $r^2 = 0.02$ achieves 80% power with $n = 399$ (ref. [15]).

These quantitative differences in required sample size could translate into large, qualitative differences in the types of neuroimaging studies considered viable in future efforts. There is a necessary trade-off between the innovativeness of a task, measure or method, and the extent to which it has been validated. Existing large-scale neuroimaging studies ($n > 1{,}000$) have selected well-validated tasks and imaging measures over new, exploratory ones, and few have attempted to characterize rare populations. Requiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation.

We agree with Marek et al.[4] that small-sample studies are important for understanding the brain bases of tasks and mental states[9–11], and for prototyping new tasks and measures. Furthermore, several current trends may further increase the viability of small-sample multivariate BWAS, including (1) new phenotypes, (2) feature-learning methods and algorithms with larger effect sizes[13], (3) models that target within-person variation in symptoms and behaviour to improve between-person predictions[2] and (4) hybrid strategies for improving prediction like meta-matching[16]. All of these have the potential to improve reliability and effect sizes, but whether they do remains to be seen.

Finally, as both Marek et al.[4] and our analyses show, very small effects will suffer from limited power, replicability and predictive utility even with sample sizes in the thousands (Fig. 1). We argue that the field should focus on discovering phenotypes and brain measures with large effect sizes. Efficient discovery entails casting a wide net in smaller studies using rigorous, unbiased methods and scaling up promising findings to larger samples[2]. There are substantial challenges ahead, including establishing broad generalizability across contexts, equity across subpopulations, and models with high neuroscientific validity and interpretability[17,18]. Addressing these challenges will require innovative new methods and measures.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05745-x.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centres that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

All analysis code used in the current study is available at GitHub (release v.1.0; https://github.com/spisakt/BWAS_comment).

1. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
2. Woo, C.-W. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: Brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
3. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry* **78**, 127–128 (2021).
4. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
5. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
6. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
7. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
8. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *Neuroimage*, **80**, 62–79 (2013).
9. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
10. Han, X. et al. Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *Neuroimage* **247**, 118844 (2022).
11. Zunhammer, M., Bingel, U. & Wager, T. D. Placebo effects on the neurologic pain signature. *JAMA Neurol.* **75**, 1321–1330 (2018).
12. Tian, Y. & Zalesky, A. Machine learning prediction of cognition from functional connectivity: are feature weights reliable? *Neuroimage* **245**, 118648 (2021).
13. Pervaiz, U., Vidaurre, D., Woolrich, M. W. & Smith, S. M. Optimising network modelling methods for fMRI. *Neuroimage* **211**, 116604 (2020).
14. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).

15. Killeen, P. R. Predict, control, and replicate to understand: how statistics can foster the fundamental goals of science. *Perspect. Behav. Sci.* **42**, 109–132 (2018).
16. He, T. et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* **25**, 795–804 (2022).
17. Wu, J. et al. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *Neuroimage* **262**, 119569 (2022).
18. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).

Acknowledgements We thank S. Marek et al. for sharing the analysis code and for the discussions in relation to our commentary. The work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; projects 'TRR289 - Treatment Expectation', ID 422744262 and 'SFB1280 - Extinction Learning', ID 316803389), R01 MH076136 and R01 EB026549.

Author contributions Conception and data analysis: T.S. Manuscript writing and revision: T.S., U.B. and T.D.W.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05745-x.

Correspondence and requests for materials should be addressed to Tamas Spisak.

nature portfolio

Corresponding author(s): Tamas Spisak

Last updated by author(s): Dec 20, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms. Preprocessed data was created software as described in: Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.
Data analysis	https://github.com/spisakt/BWAS_comment v1.0 Dependencies: python 3.10.8 numpy 1.23.5 pandas 1.5.2 scikit-learn 1.2.0 joblib 1.2.0 mlxtend 0.21.0 seaborn 0.12.1 matplotlib 3.6.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
All derivative data, (including raw material for figures) is freely available at https://github.com/spisakt/BWAS_comment

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The Human Connectome Project involved 656 females and 550 males. Findings apply to both males and females. Sex and gender differences are out of the scope of the Matters Arising and were not considered in the analysis. For more information, refer to van Essen et al., 2013.
Population characteristics	Refer to van Essen et al., 2013.
Recruitment	As described in van Essen et al., 2013.
Ethics oversight	As described in van Essen et al., 2013.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the replicability of functional connectivity-based brain-wise association studies.
Research sample	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project (van Essen et al., 2013), based on a non-representative sample of young adults (656 females, 550 males, mean+-sd age: 28.9+-3.57).
Sampling strategy	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project. For more information, please refer to van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022.
Data collection	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Timing	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Data exclusions	We excluded participants with no MRI images available, and for each analysis, participants with missing data about the target phenotype.
Non-participation	203 out of the 1206 HCP participants didn't have MRI data. Exclusion die to missing phenotype data varied across analyses (less than 20 in all cases)
Randomization	N/A (observational study)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	resting state fMRI, anatomical MRI
Design specifications	4, 15min, eyes open (see van Essen et al., 2013)
Behavioral performance measures	As described in van Essen et al., 2013.

Acquisition

Imaging type(s)	functional, anatomical
Field strength	3
Sequence & imaging parameters	As described in van Essen et al., 2013.
Area of acquisition	brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	As described in: Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.
Normalization	As described in: van Essen et al., 2013.
Normalization template	As described in: van Essen et al., 2013.
Noise and artifact removal	As described in: van Essen et al., 2013.
Volume censoring	As described in: van Essen et al., 2013.

Statistical modeling & inference

Model type and settings	Multivariate predictive modelling, PCA+SVR and Ridge, cross-validation (see Supplementary Methods for details).
Effect(s) tested	Predictive performance (Pearson's r)

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	N/A (predictive modelling)
Correction	N/A (predictive modelling)

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	partial correlation
Multivariate modeling and predictive analysis	Targets: age, cognitive ability, episodic memory, fluid intelligence, cognitive flexibility, inhibition. Model: PCA +SVR (reproduction of Marek et al.'s model), Ridge regression (with the default hyperparameter value 1). The PCA+SVR model involved dimensionality reduction. Features partial correlation values across 100 ICA-based regions, training metric: mean squared error, evaluation metric, mean squared error and Pearson correlation.

Reply to: Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05746-w

Published online: 8 March 2023

Open access

Check for updates

Brenden Tervo-Clemmens[1,22]✉, Scott Marek[2,3,22]✉, Roselyne J. Chauvin[4], Andrew N. Van[4,5], Benjamin P. Kay[4], Timothy O. Laumann[3], Wesley K. Thompson[6], Thomas E. Nichols[7], B. T. Thomas Yeo[8,9,10,11,12,13], Deanna M. Barch[3,14], Beatriz Luna[15,16], Damien A. Fair[17,18,19,23]✉ & Nico U. F. Dosenbach[2,4,5,14,20,21,23]✉

REPLYING TO: T. Spisak et al. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023)

In our previous study[1], we documented the effect of sample size on the reproducibility of brain-wide association studies (BWAS) that aim to cross-sectionally relate individual differences in human brain structure (cortical thickness) or function (resting-state functional connectivity (RSFC)) to cognitive or mental health phenotypes. Applying univariate and multivariate methods (for example, support vector regression (SVR)) to three large-scale neuroimaging datasets (total $n \approx 50{,}000$), we found that overall BWAS reproducibility was low for $n < 1{,}000$, due to smaller than expected effect sizes. When samples and true effects are small, sampling variability, and/or overfitting can generate 'statistically significant' associations that are likely to be reported due to publication bias, but are not reproducible[2–5], and we therefore suggested that BWAS should build on recent precedents[6,7] and continue to aim for samples in the thousands. In the accompanying Comment, Spisak et al.[8] agree that larger BWAS are better[5,9], but argue that "multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases" ($n = 75$–500); this suggestion is made on the basis of analyses of a selected subset of multivariate cognition/RSFC associations with larger effect sizes, using their preferred method (ridge regression with partial correlations) in a demographically more homogeneous, single-site/scanner sample (Human Connectome Project (HCP), $n = 1{,}200$, aged 22–35 years).

There is no disagreement that a minority of BWAS effects can replicate in smaller samples, as shown with our original methods[1]). Using the exact methodology (including cross-validation) and code of Spisak et al.[8] to repeat 64 multivariate BWAS in the 21-site, larger and more diverse Adolescent Brain Cognitive Development Study (ABCD, $n = 11{,}874$, aged 9–11 years), we found that 31% replicated at $n = 1{,}000$, dropping to 14% at $n = 500$ and none at $n = 75$. Contrary to the claims of Spisak et al.[8], replication failure was the outcome in most cases when applied to this larger, more diverse dataset. Basing general BWAS sample size recommendations on the largest effects has at least two fundamental flaws: (1) failing to detect other true effects (for example, reducing the sample size from $n = 1{,}000$ to $n = 500$ leads to a 55% false-negative rate), therefore restricting BWAS scope, and (2) inflation of reported effects[3,10–12]. Thus, regardless of the method, associations based on small samples can remain distorted and lack generalizability until confirmed in large, diverse, independent samples.

We always test for BWAS replication with null models (using permutation tests) of out-of-sample estimates to ensure that our reported reproducibility is unaffected by in-sample overfitting. Nonetheless, Spisak et al.[8] argue against plotting inflated in-sample estimates[1,10] on the y axis, and out-of-sample values on the x axis, as we did (Fig. 1a). Instead, they propose plotting cross-validated associations from an initial, discovery sample (Fig. 1b (y axis)) against split-half out-of-sample associations (x axis). However, cross-validation—just like split-half validation—estimates out-of-sample, and not in-sample, effect sizes[13]. The in-sample associations[1,10] for the method of Spisak et al.[8] (Fig. 1b), that is, from data in the sample used to develop the model, show the same degree of overfitting (Fig. 1a versus Fig. 1b). The plot of Spisak et al.[8] (Fig. 1c) simply adds an additional out-of-sample test (cross-validation before split half), and therefore demonstrates the close correspondence between two different methods for out-of-sample effect estimation[14]. Analogously, we can replace the cross-validation step in the code of Spisak et al.[8] with split-half validation (our original out-of-sample test), obtaining split-half effects in the first half of the sample, and then comparing them to the split-half estimates from the full sample (Fig. 1d). The strong correspondences between cross-validation followed by split-half (Spisak et al. method[8]; Fig. 1c) and repeated split-half validation (Fig. 1d) are guaranteed by plotting out-of-sample estimates (from the same dataset) against one another. Here, plotting cross-validated discovery sample estimates on the y axis (Fig. 1c,d) provides no additional information beyond the x axis out-of-sample values. The critically important out-of-sample predictions, required for reporting multivariate results[1], generated using the method of Spisak et al.[8] and our method are nearly identical (Fig. 1e).

As Spisak et al.[8] highlight, cross-validation of some type is considered to be standard practice[10], and yet the distribution of out-of-sample

[1]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Radiology, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA. [4]Department of Neurology, Washington University School of Medicine, St Louis, MO, USA. [5]Department of Biomedical Engineering, Washington University in St Louis, St Louis, MO, USA. [6]Division of Biostatistics, University of California San Diego, La Jolla, CA, USA. [7]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [8]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore. [9]Centre for Sleep and Cognition, National University of Singapore, Singapore, Singapore. [10]Centre for Translational MR Research, National University of Singapore, Singapore, Singapore. [11]N.1 Institute for Health, Institute for Digital Medicine, National University of Singapore, Singapore, Singapore. [12]Integrative Sciences and Engineering Programme, National University of Singapore, Singapore, Singapore. [13]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA. [14]Department of Psychological and Brain Sciences, Washington University in St Louis, St Louis, MO, USA. [15]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA. [16]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. [17]Masonic Institute for the Developing Brain, University of Minnesota Medical School, Minneapolis, MN, USA. [18]Department of Pediatrics, University of Minnesota Medical School, Minneapolis, MN, USA. [19]Institute of Child Development, University of Minnesota Medical School, Minneapolis, MN, USA. [20]Program in Occupational Therapy, Washington University School of Medicine, St Louis, MO, USA. [21]Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA. [22]These authors contributed equally: Brenden Tervo-Clemmens, Scott Marek. [23]These authors jointly supervised this work: Damien A. Fair, Nico U. F. Dosenbach. ✉e-mail: btervo-clemmens@mgh.harvard.edu; smarek@wustl.edu; faird@umn.edu; ndosenbach@wustl.edu

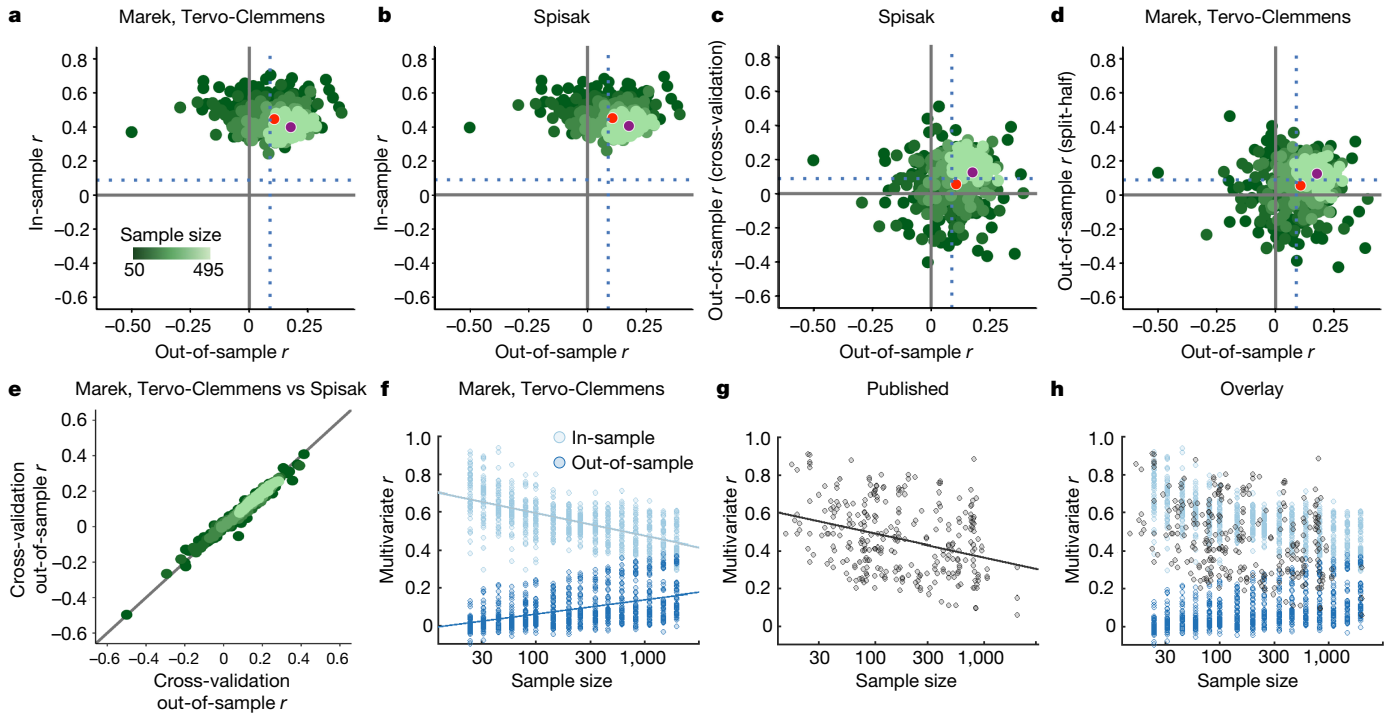

Fig. 1 | In-sample versus out-of-sample effect estimates in multivariate BWAS. a–**e**, Methods comparison between our previous study[1] (split-half) and Spisak et al.[8] (cross-validation followed by split-half). 'Marek, Tervo-Clemmens' and 'Spisak' refer to the methodolgies described in ref. [1] and ref. [8], respectively. For **a**–**e**, HCP 1200 Release (full correlation) data were used to predict age-adjusted total cognitive ability. Analysis code and visualizations (*x*,*y* scaling; colours) are the same as in Spisak et al.[8]. The *x* axes in **a**–**e** always display the split-half out-of-sample effect estimates from the second (replication) half of the data (correlation between true scores and predicted scores; as in Spisak et al.[8] and in our previous study[1]; Supplementary Methods). **a**, In-sample (training correlation; *y* axis) as a function of out-of-sample associations (plot convention in our previous study[1]). **b**, Matched comparison of the true in-sample association (training correlations, mean across folds; *y* axis) in the method proposed by Spisak et al.[8]. **c**, The proposed correction by Spisak et al.[8] that inserts an additional cross-validation step to evaluate the first half of data, which by definition makes this an out-of-sample association (*y* axis). **d**, Replacing the cross-validation step from Spisak et al.[8] with a split-half validation provides a different (compared with **c**) out-of-sample association of the first half of the total data (that is, each of the first stage split halves is one-quarter of the total data; *y* axis). The appropriate and direct comparison of in-sample associations between Spisak et al.[8] and our previous study[1] is comparing **b** to **a**, rather than **c** to **a**. The Spisak et al. method[8] (cross-validation followed by split-half validation) does not reduce in-sample overfitting (**b**) but, instead, adds an additional out-of-sample evaluation (**c**), which is nearly identical to split-half validation twice in a row (**d**), and makes it clear why the out-of-sample performance of these two methods is likewise nearly identical. **e**, Correspondence between out-of-sample associations (to the left-out half) from the additional cross-validation step proposed by Spisak et al.[8] (mean across folds; *y* axis) and the original split-half validation from our previous study[1] (*x* axis). The identity line is shown in black. **f**, In-sample (*r*; light blue) and out-of-sample (*r*; dark blue) associations as a function of sample size. Data are from figure 4a–d of ref. [1]. **g**, Published literature review of multivariate *r* (*y* axis) as a function of sample size (data from ref. [15]) displayed with permission. For **f** and **g**, best fit lines are displayed in $\log_{10}$ space. **h**, Overlap of **f** and **g**.

associations (Fig. 1f (dark blue)) does not match published multivariate BWAS results (Fig. 1g), which have largely ranged from $r = 0.25$ to 0.9, decreasing with increasing sample size[10,15,16]. Instead, published effects more closely follow the distribution of in-sample associations (Fig. 1h). This observation suggests that, in addition to small samples, structural problems in academic research (for example, non-representative samples, publication bias, misuse of cross-validation and unintended overfitting) have contributed to the publication of inflated effects[12,17,18]. A recent biomarker challenge[5] showed that cross-validation results continued to improve with the amount of time researchers spent with the data, and the models with the best cross-validation results performed worse on never-seen held-back data. Thus, cross-validation alone has proven to be insufficient and must be combined with the increased generalizability of large, diverse datasets and independent out-of-sample evaluation in new, never before seen data[5,10].

The use of additional cross-validation in the discovery sample by Spisak et al.[8] does not affect out-of-sample prediction accuracies (Fig. 1e). However, by using partial correlations and ridge regression on HCP data, they were able to generate higher out-of-sample prediction accuracies than our original results in ABCD (Fig. 2a). The five variables they selected are strongly correlated[19] cognitive measures from the NIH Toolbox (mean $r = 0.37$; compare with the correlation strength for height versus weight $r = 0.44$)[20] and age (not a complex behavioural phenotype), unrepresentative of BWAS as a whole (Fig. 2b (colour versus grey lines)). As the HCP is the relatively smallest and most homogeneous dataset, we applied the exact method and code of Spisak et al.[8] to the ABCD data (Fig. 2c and Supplementary Table 2). At $n = 1{,}000$ (training; $n = 2{,}000$ total), only 31% of BWAS (44% RSFC, 19% cortical thickness) were replicable (Fig. 2d; defined as in Spisak et al.[8]; Supplementary Information). Expanding BWAS scope beyond broad cognitive abilities towards complex mental health outcomes therefore requires $n > 1{,}000$ (Fig. 2b–d). The absolute largest BWAS (cognitive ability: RSFC, green) reached replicability only using $n = 400$ ($n = 200$ train; $n = 200$ test) with an approximate 40% decrease in out-of-sample prediction accuracies from HCP to ABCD (Fig. 2e (lighter green, left versus right)). The methods of Spisak et al.[8] and our previous study[1] returned equivalent out-of-sample reproducibility for this BWAS (cognitive ability: RSFC) in the larger, more diverse ABCD data (Fig. 2e (right, dark versus light green)). Thus, the smaller sample sizes (Fig. 2b,c) that are required for out-of-sample reproducibility (Fig. 2e) reported by Spisak et al.[8] in the HCP data did not generalize to the larger ABCD dataset. See also our previous study[1] for a broader discussion of convergent evidence across HCP and ABCD datasets.

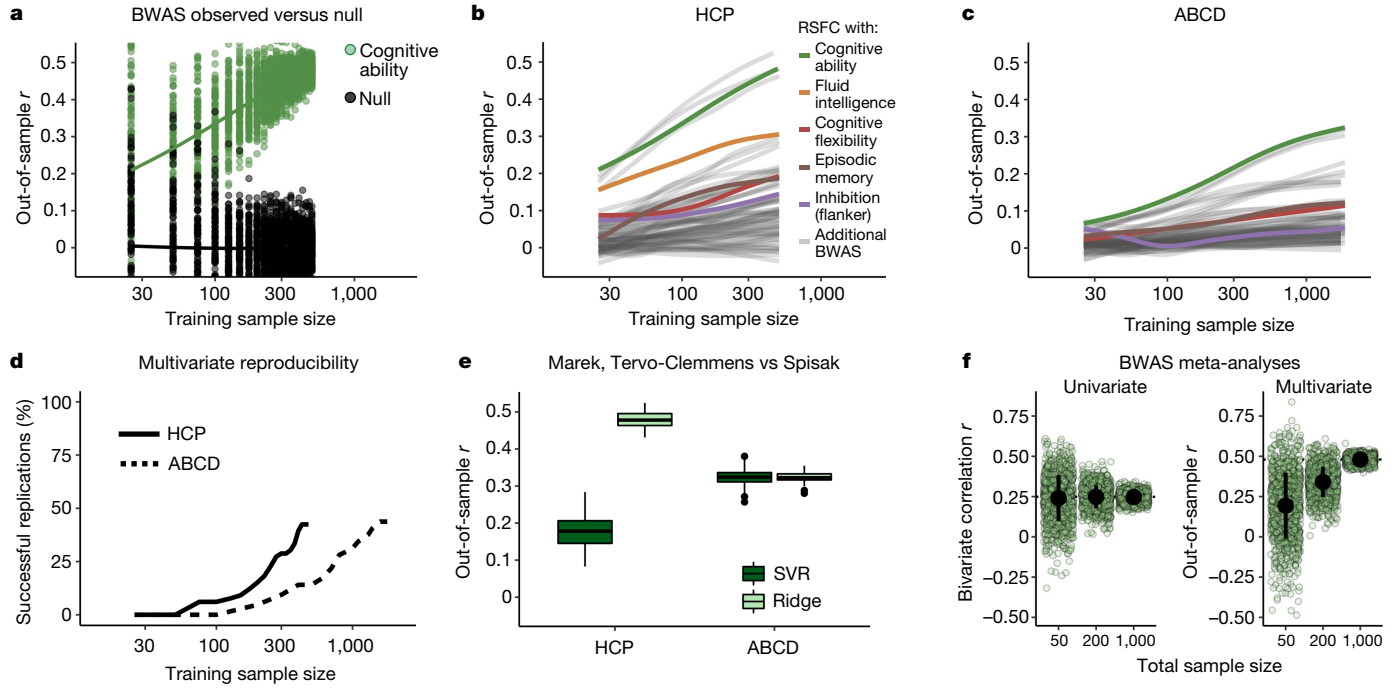

Fig. 2 | BWAS reproducibility, scope and prediction accuracy using the method of Spisak et al. a, Example bootstrapped BWAS of total cognitive ability (green) and null distribution (black) (*y* axis), as a function of sample size (*x* axis) from the suggested method of Spisak et al.[8] (RSFC by partial correlation; prediction by ridge regression) in the HCP dataset ($n = 1{,}200$, 1 site, 1 scanner, 60 min RSFC/participant, 76% white). Sample sizes were $\log_{10}$-transformed for visualization. **b**, Out-of-sample correlation (between true scores and predicted scores) from ridge regression (*y* axis; code from Spisak et al.[8]) as a function of training sample size (*x* axis, $\log_{10}$ scaling) for 33 cognitive and mental health phenotypes (Supplementary Information) in the HCP dataset. Each line displays a smoothed fit estimate (through penalized splines in general additive models) for a brain (RSFC (partial correlations, as proposed by Spisak et al.[8]), cortical thickness) phenotype pair (66 total) that has 100 bootstrapped iterations from sample sizes of 25 to 500 (inclusive) in increments of 25 (20 total bins). Sample sizes were $\log_{10}$-transformed (for visualization) before general additive model fitting. **c**, The same as in **b**, but in the ABCD dataset ($n = 11{,}874$, 21 sites, 3 scanner manufacturers, 20 min RSFC/participant, 56% white) using 32 cognitive and mental health phenotypes at sample sizes of 25, 50, 75 and from 100 to 1,900 (inclusive) in increments of 100 (22 total bins). **d**, The percentage of brain–phenotype pairs (BWAS) from **b** and **c** with significant replication on the basis of the method of Spisak et al.[8] (Supplementary Information). **e**, Comparison of our original method in our previous study[1] and the method proposed by Spisak et al.[8] at the full split-half sample size of HCP (left) and ABCD (right). Out-of-sample correlations (RSFC with total cognitive ability, *y* axis) for the method used in our previous study[1] (dark green; RSFC by correlation, PCA, SVR) and by Spisak et al.[8] (light green; RSFC by partial correlation, ridge regression). Repeating the method proposed by Spisak et al.[8] in ABCD (right) and comparing this to the method used in our previous study[1] results in a very similar out-of-sample *r*. **f**, Simulated individual studies (light green circles; $n = 1{,}000$ per sample size) and meta-analytic estimates (black dot, ±1 s.d.) using the method of Spisak et al.[8] (partial correlations in the HCP dataset) for the largest univariate association (left; *y* axis, bivariate correlation) and multivariate association (right; *y* axis, out-of-sample correlation) for total cognitive ability versus RSFC, as a function of total sample size (*x* axis; bivariate correlation for sample sizes of 50, 200 and 1,000, and multivariate sum of train and test samples, each 25, 100 and 500). For univariate approaches, studies of any sample size, when appropriately aggregated to a large total sample size, can correctly estimate the true effect size. However, for multivariate approaches, even when aggregating across 1,000 independent studies, studies with a small sample size produce prediction accuracies that are downwardly biased relative to large sample studies, highlighting the need for large samples in multivariate analyses.

Notably, the objections of Spisak et al.[8] raise additional reasons to stop the use of smaller samples in BWAS that were not highlighted in our original article. Multivariate BWAS prediction accuracies—absent overfitting—are systematically suppressed in smaller samples[5,9,21], as prediction accuracy scales with increasing sample size[1,9]. Thus, the claim that "cross-validated discovery effect-size estimates are unbiased" does not account for out-of-dataset generalizability and downward bias. In principle, if unintended overfitting and publication bias could be fully eliminated, meta-analyses of small-sample univariate BWAS would return the correct association strengths (Fig. 2f (left)). However, meta-analyses of small multivariate BWAS would always be downwardly biased (Fig. 2f (right)). If we are interested in maximizing prediction accuracy, essential for clinical implementation of BWAS[22], large samples and advancements in imaging and phenotypic measurements[1] are necessary.

Repeatedly subsampling the same dataset, as Spisak et al.[8] and we have done, overestimates reproducibility compared with testing on a truly new, diverse dataset. Just as in genomics[23], BWAS generalization failures have been highlighted[5,24]. For example, BWAS models trained on white Americans transferred poorly to African Americans and vice versa (within dataset)[24]. Historically, BWAS samples have lacked diversity, neglecting marginalized and under-represented minorities[25]. Large studies with more diverse samples and data aggregation efforts can improve BWAS generalizability and reduce scientific biases contributing to massive health inequities[26,27].

Spisak et al.[8] worry that "[r]equiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation". We appreciate the concern that rarer populations may never be investigated with BWAS. Yet, there are many non-BWAS brain–behaviour study designs (fMRI ≠ BWAS) focused on within-patient effects, repeated-sampling and signal-to-noise-ratio improvements that have proven fruitful down to $n = 1$ (ref. [28]). By contrast, the strength of multivariate BWAS lies in leveraging large cross-sectional samples to investigate population-level questions. Sample size requirements should be based on expected effect sizes and real-world impact, and not resource availability. Through large-scale collaboration and clear standards on data sharing, GWAS has reached sample sizes in the millions[29–31], pushing genomics towards new horizons. Similarly, BWAS

analyses of the future will not be limited to statistical replication of the same few strongest effects in small homogeneous populations, but also have broad scope, maximum prediction accuracy and excellent generalizability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05746-w.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Participant-level data from all datasets (ABCD and HCP) are openly available pursuant to individual, consortium-level data access rules. The ABCD data repository grows and changes over time (https://nda.nih.gov/abcd). The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0 (https://doi.org/10.15154/1503209). Data were provided, in part, by the HCP, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the HCP (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

Manuscript analysis code specific to this study is available at GitHub (https://gitlab.com/DosenbachGreene/bwas_response). Code for processing ABCD data is provided at GitHub (https://github.com/DCAN-Labs/abcd-hcp-pipeline). MRI data analysis code is provided at GitHub (https://github.com/ABCD-STUDY/nda-abcd-collection-3165). FIRMM software is available online (https://firmm.readthedocs.io/en/latest/release_notes/). The ABCD Study used v.3.0.14.

1. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
2. Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *J. Res. Pers.* **47**, 609–612 (2013).
3. Button, K. S. et al. Confidence and precision increase with high statistical power. *Nat. Rev. Neurosci.* **14**, 585–586 (2013).
4. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).
5. Traut, N. et al. Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *Neuroimage* **255**, 119171 (2022).
6. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
7. Littlejohns, T. J. et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
8. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023).
9. Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D. & Ritter, K. Performance reserves in brain-imaging-based phenotype prediction. Preprint at https://doi.org/10.1101/2022.02.23.481601 (2022).
10. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
11. Poldrack, R. A. The costs of reproducibility. *Neuron* **101**, 11–14 (2019).
12. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
13. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. IJCAI 95* (ed. Mellish, C. S.) 1137–1143 (Morgan Kaufman, 1995).
14. Scheinost, D. et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* **193**, 35–45 (2019).
15. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* **88**, 818–828 (2020).
16. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
17. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
18. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 108–120 (2019).
19. Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168 (2019).
20. Meyer, G. J. et al. Psychological testing and psychological assessment: a review of evidence and issues. *Am. Psychol.* **56**, 128–165 (2001).
21. He, T. et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* **206**, 116276 (2020).
22. Leptak, C. et al. What evidence do we need for biomarker qualification? *Sci. Transl. Med.* **9**, eaal4599 (2017).
23. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
24. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).
25. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
26. Bailey, Z. D. et al. Structural racism and health inequities in the USA: evidence and interventions. *Lancet* **389**, 1453–1463 (2017).
27. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
28. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain-behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
29. Levey, D. F et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nature* **24**, 954–963 (2021).
30. Muggleton, N. et al. The association between gambling and financial, social and health outcomes in big financial data. *Nat. Hum. Behav.* **5**, 319–326 (2021).
31. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

Acknowledgements Data used in the preparation of this Article were, in part, obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9–10 years and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093 and U01DA041025. A full list of supporters is available online (https://abcdstudy.org/federal-partners.html). A listing of participating sites and a complete listing of the study investigators is available online (https://abcdstudy.org/scientists/workgroups/). ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or the writing of this report. This Article reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. Data were provided, in part, by the HCP, WU-Minn Consortium (U54 MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This work used the storage and computational resources provided by the Masonic Institute for the Developing Brain (MIDB), the Neuroimaging Genomics Data Resource (NGDR) and the Minnesota Supercomputing Institute (MSI). The NGDR is supported by the University of Minnesota Informatics Institute through the MnDRIVE initiative in coordination with the College of Liberal Arts, Medical School and College of Education and Human Development at the University of Minnesota. This work used the storage and computational resources provided by the Daenerys Neuroimaging Community Computing Resource (NCCR). The Daenerys NCCR is supported by the McDonnell Center for Systems Neuroscience at Washington University, the Intellectual and Developmental Disabilities Research Center (IDDRC; P50 HD103525) at Washington University School of Medicine and the Institute of Clinical and Translational Sciences (ICTS; UL1 TR002345) at Washington University School of Medicine. This work was supported by NIH grants MH121518 (to S.M.), NS090978 (to B.P.K.), MH129616 (to T.O.L.), 1RF1MH120025-01A1 (to W.K.T), MH080243 (to B.L.), MH067924 (to B.L.), DA041148 (to D.A.F.), DA04112 (to D.A.F.), MH115357 (to D.A.F.), MH096773 (to D.A.F. and N.U.F.D.), MH122066 (to D.A.F. and N.U.F.D.), MH121276 (to D.A.F. and N.U.F.D.), MH124567 (to D.A.F. and N.U.F.D.), NS088590 (to N.U.F.D.), and the Andrew Mellon Predoctoral Fellowship (to B.T.-C.), the Staunton Farm Foundation (to B.L.), the Lynne and Andrew Redleaf Foundation (to D.A.F.) and the Kiwanis Neuroscience Research Foundation (to N.U.F.D.).

Author contributions Conception: B.T.-C., S.M., D.A.F. and N.U.F.D. Design: B.T.-C., S.M., R.J.C., D.A.F. and N.U.F.D. Data acquisition, analysis and interpretation: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., W.K.T., T.E.N., B.T.T.Y., D.A.F. and N.U.F.D. Manuscript writing and revising: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., T.O.L., W.K.T., T.E.N., B.T.T.Y., D.M.B., B.L., D.A.F. and N.U.F.D. We note that the reply author list differs from the original paper in number and in order to accurately reflect its more focused scope compared with the original work.

Competing interests D.A.F. and N.U.F.D. have a financial interest in Turing Medical and may financially benefit if the company is successful in marketing FIRMM motion monitoring software products. A.N.V., D.A.F. and N.U.F.D. may receive royalty income based on FIRMM technology developed at Washington University School of Medicine and Oregon Health and Sciences University and licensed to Turing Medical. D.A.F. and N.U.F.D. are co-founders of Turing Medical. These potential conflicts of interest have been reviewed and are managed by Washington University School of Medicine, Oregon Health and Sciences University and the University of Minnesota.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05746-w.

Correspondence and requests for materials should be addressed to Brenden Tervo-Clemmens, Scott Marek, Damien A. Fair or Nico U. F. Dosenbach.

nature portfolio

Corresponding author(s): Scott Marek, Brenden Tervo-Clemmens, Nico Dosenbach

Last updated by author(s): Dec 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. Neuroimaging and behavioral data were from existing, open-source datasets (ABCD, UKB, HCP) whose acquisition's are presented in detail in previous work. The ABCD Study data were collected between 2016-2018. The HCP data were collected between 2010-2016.
Data analysis	MRI data analysis code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165 ABCD and UKB MRI data processing code can be found here https://github.com/DCAN-Labs/abcd-hcp-pipeline Manuscript analysis code can be found here https://gitlab.com/DosenbachGreene/bwas_response FIRMM software: https://firmm.readthedocs.io/en/latest/release_notes/. ABCD uses version 3.0.14.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Participant level data from all datasets (ABCD & HCP) is openly available pursuant to individual, consortia-level data access rules. The ABCD data repository grows and changes over time. The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0, DOI 10.15154/1503209.

Data were provided, in part, by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
No new data were collected for this manuscript. Across the ABCD, and HCP we downloaded data between 01/2019 - 10/2021. We did not use any specific software for downloading the data. For details on data collection in ABCD (baseline data), see Casey et al., 2018; in HCP (1200 release) see Van Essen et al., 2013).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the magnitude and reproducibility of cross-sectional associations between neuroimaging measures and psychological/psychiatric phenotypes.
Research sample	Our main focus was to replicate work from Spisak et al in both the dataset used in their paper (HCP) and to further test generalizability of their models. Therefore, we also tested their models on the larger ABCD dataset.
Sampling strategy	All samples were recruited from the community (ABCD & HCP from the USA). Individual samples (ABCD, HCP) used unique sample size calculations and sampling strategies which are discussed in prior work with these open source datasets (Casey et al., 2018, Van Essen et al., 2013, respectively).
Data collection	All data were from existing data repositories and were downloaded between 01/2019 - 10/2021. Data used in the manuscript were from existing large consortia datasets (ABCD: see Casey et al., 2018 & Barch et al., 2018; HCP: We used data from the 1200 subjects data release (van Essen et al., 2013). Because we did not personally collect any of the data used in this manuscript, all data were from existing data repositories and researchers were therefore not blind to the source of the data.
Timing	ABCD: see Casey et al., 2018 HCP: see van Essen et al., 2013
Data exclusions	In ABCD, we used strict inclusion criteria with regard to head motion. Specifically, inclusion criteria for the current project consisted of at least 600 frames (8 minutes) of low-motion (filtered FD<0.08) resting state functional connectivity data. Our final dataset consisted of data from a total of N=3,928 youth across the discovery (N=1,964) and replication (N=1,964) sets. The final discovery and replication sets did not differ in mean FD (ΔM=0.002 , t=0.60, p=0.55) or total frames included (ΔM=6.4 , t =0.94, p=0.35). The subject lists for ARMS samples and our associated matrices will be released in the ABCD-BIDS Community Collection (ABCD collection 3165) for community use. For HCP data, we used similar data quantity inclusion, as well as an FD < 0.20 (unfiltered FD). This resulted in the inclusion of N=900 individuals (N=877 across all NIH Toolbox subscales).
Non-participation	N/A
Randomization	All three samples were observational studies and no randomization was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	See above.
Ethics oversight	The ABCD Study obtained centralized institutional review board approval from the University of California, San Diego, and each of the 21 study sites obtained local institutional review board approval. Ethical regulations were followed during data collection and analysis. Parents or caregivers provided written informed consent, and children gave written assent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	resting-state fMRI, task-based fMRI; structural (cortical thickness) MRI
Design specifications	ABCD resting state: 4, 5 min runs, eyes open HCP resting state: 4, 15 min runs, eyes open
Behavioral performance measures	Primary analyses use cognitive assessments from the NIH Toolbox and psychopathology assessment Child Behavior Checklist (see manuscript for individual subscales, total of 41) included in standard data releases and discussed in detail perviously (Barch et al., 2018)

Acquisition

Imaging type(s)	Resting-state fMRI, task-fMRI, structural (cortical thickness) MRI
Field strength	3 Tesla
Sequence & imaging parameters	Primary analyses use open-source distributed fMRI and MR data that adhere to consortia guidelines (see Casey et al., 2018 and Van Essen et al., 2013, for ABCD and HCP, respectively).
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Preprocessing of ABCD was done using a suite of tools. All code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165. Individual datasets (ABCD, HCP) and individual study sites (e.g., ABCD site 1 versus site 2) used unique sequence and imaging parameters which are discussed in prior work introducing these open-source datasets.
Normalization	1) PreFreesurfer normalizes anatomical data. This normalization entails brain extraction, denoising, and then bias field correction on anatomical T1 and/or T2 weighted data. The ABCD-HCP pipeline includes two additional modifications to improve output image quality. ANTs 65 DenoiseImage models scanner noise as a Rician distribution and attempts to remove such noise from the T1 and T2 anatomical images. Additionally, ANTs N4BiasFieldCorrection attempts to smooth relative image histograms in different parts of the brain and improves bias field correction. 2) FreeSurfer 1 constructs cortical surfaces from the normalized anatomical data. This stage performs anatomical segmentation, white/grey and grey/CSF cortical surface construction, and surface registration to a standard surface template. Surfaces are refined using the T2 weighted anatomical data. Mid-thickness surfaces, which represent the average of white/grey and grey/CSF surfaces, are generated here. 3) PostFreesurfer converts prior outputs into an HCP-compatible format (i.e. CIFTIs) and transforms the volumes to a standard volume template space using ANTs nonlinear registration, and the surfaces to the standard surface

space via spherical registration.

Normalization template

The "Vol" stage corrects for functional distortions via reverse-phase encoding spin-echo images. All resting state runs underwent intensity normalization to a whole brain mode value of 1000, within run correction for head movement, and functional data registration to the standard template (MNI). Atlas transformation was computed by registering the mean intensity image from each BOLD session to the high resolution T1 image, and then applying the anatomical registration to the BOLD image. This atlas transformation, mean field distortion correction, and resampling to 3-mm isotropic atlas space were combined into a single interpolation using FSL's 66 applywarp tool. The "Surf" stage projects the normalized functional data onto the template surfaces.

Noise and artifact removal

Additional BOLD preprocessing steps were executed to reduce spurious variance unlikely to reflect neuronal activity 46. First, a respiratory filter was used to improve FD estimates calculated in the volume ("vol") stage68. Second, temporal masks were created to flag motion-contaminated frames using the improved FD estimates 63. Frames with a filtered FD>0.3mm were flagged as motion-contaminated for nuisance regression only. After computing the temporal masks for high motion frame censoring, the data were processed with the following steps: (i) demeaning and detrending, (ii) interpolation across censored frames using least squares spectral estimation of the values at censored frames so that continuous data can be (iii) denoised via a GLM with whole brain, ventricular, and white matter signal regressors, as well as their derivatives. Denoised data were then passed through (iv) a band-pass filter (0.008 Hz<f<0.10 Hz) without re-introducing nuisance signals 69 or contaminating frames near high motion frames.

Volume censoring

Yes, ABCD data were censored at a filtered frame-wise displacement of < 0.08mm and HCP data were filtered using a non-filtered framewise displacement of <0.20mm.

Statistical modeling & inference

Model type and settings

Mass univariate and multivariate (support vector regression, canonical correlation analysis). Multiple parameterizations of each of these models were explored with the stated goal being to determine field-wide reproducibility in brain-phenotype association studies (see manuscript).

Effect(s) tested

As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple scales and combinations of effects were examined. Owing to the cross-sectional, nature of these studies, all effects are between-person associations.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)

Parcel-level and network-level analyses utilized the field-standard Gordon et al., 2016, Cerebral Cortex, and Seitzman et al., 2020, NeuroImage. Vertex-wise and voxel-wise data were extracted from Ciftis.

Statistic type for inference (See Eklund et al. 2016)

Multiple levels of neuroanatomical scale were used, including voxels, regions of interest, and networks.

Correction

As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple levels of significance values and correction were used, ranging from uncorrected to bonferroni (FWER) correction.

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Pearson correlation

Multivariate modeling and predictive analysis

Two supervised regression models were used: a Ridge Regression model (α = 1.0), as proposed by Spisak et al. and a combined Principal Component Analysis (PCA) and Support Vector Regression (SVR) model, whereby half of the principal components (retaining 50% of the variance) generated from the PCA were passed as features into the SVR, as in the original work by Marek, Tervo-Clemmens et al. Both models were implemented using scikit-learn 5 in Python 3.

For both HCP and ABCD datasets, both methods (ridge regression; PCA+SVR) and using three different neuroimaging feature sets (RSFC: full correlation, partial correlation; cortical thickness), the same analyses were conducted using code directly from Spisak et al. (https://gitlab.com/DosenbachGreene/bwas_response). For each behavioural phenotype and neuroimaging feature set combination, in each dataset, a complete cases sub-dataset was compiled, removing participants with missing behavioural phenotypes or neuroimaging data. For each of these complete cases (per Spisak et al.) neuroimaging feature set behavioural phenotype sub-datasets, 100 bootstraps were run for each model. Within each bootstrap, the sub-dataset was equally and randomly split into a discovery and replication set based on a given sample size. Here, sample size is defined as the size of a sole discovery/training set (identical in size to the replication set), such that given a sample size n, the total number of participants/samples of the combined discovery and replication sets is 2n.

Following Spisak et al. (method and code), the discovery set was divided again into 10 cross-validation folds. However, unlike the nested cross-validation which was explored in our original manuscript and shown to not substantively change results (Marek, Tervo-Clemmens et al. 1: Supplemental Fig. S11, S12), this procedure

utilised by Spisak et al., and repeated here, did not use the additional cross-validation step for hyperparameter tuning. Rather an additional out-of-sample test was applied to the discovery dataset. TThe analyses and Figures (Fig. 1, 2) in this work use combinations of Spisak et al.'s methodological suggestions and those from our original work to replicate, expand, and clarify Spisak et al.'s Matters Arising commentary and to provide a more comprehensive perspective on out-of-sample multivariate BWAS effects. Rationale and additional details for specific analyses are provided in the relevant "Main Text" and "Figure Captions". In all cases, out-of-sample associations were evaluated as the correlation between the predicted phenotype score and the true score in the out-of-sample data. In-sample (training) associations were evaluated as the correlation between the true score and the predicted score from the model developed in the discovery set (that is, the data in the sample used to develop the model (Fig. 1).

Successful out-of-sample replication was defined as in Spisak et al.: 80% of bootstrapped iterations for a given behavioural phenotype-brain feature set ("BWAS") that were significant (via permutation test) in the first cross-validation test are significant in the second, split half test. We note this definition of replication by Spisak et al. thus does not consider all bootstrap iterations (n = 100) run when determining replication success/failure. That is, the denominator of a replication percentage is set by the number of bootstrap iterations that are significant in the first cross-validation test. Therefore, to ensure this measure of 80% replication represented a true percentage, replication here also required that more than one bootstrap iteration (out of the total 100) replicated (as defined above). Without this criteria, the impact of sampling variability and the performance of a single bootstrap iteration ensured that a small number of BWAS would appear to intermittently have replication successes followed by replication failure for the very smallest sample sizes. Reproducibility estimates following Spisak et al. guidelines were highly consistent with those from our original work.