CellPress

Neuron

Spotlight

Brain-behavior correlations: Two paths toward reliability

Caterina Gratton,[1,2,*] Steven M. Nelson,[3,4] and Evan M. Gordon[5]
[1]Department of Psychology, Northwestern University, Evanston, IL, USA
[2]Department of Neurology, Northwestern University, Evanston, IL, USA
[3]Department of Pediatrics, University of Minnesota, Minneapolis, MN, USA
[4]Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN, USA
[5]Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA
*Correspondence: caterina.gratton@gmail.com
https://doi.org/10.1016/j.neuron.2022.04.018

In a recent issue of *Nature*, Marek et al. (2022) demonstrate that cross-sectional brain-behavior correlations are often small and unreliable without large samples. This observation pushes human neuroscience toward study designs that either maximize sample sizes to detect small effects or maximize effect sizes using focused investigations.

Neuroscience is centered on understanding how the nervous system relates to behavior and using this knowledge to guide clinical care in neurological and psychiatric disorders. One common approach used to attack this core question is to examine brain-behavior associations across people—that is, to examine how variation in some brain measure across individuals (e.g., dorsolateral prefrontal cortex activity) relates to variation in some behavior (e.g., depression symptoms). This approach contrasts with study designs that examine links between brain and behavior using theoretically focused, controlled, within-subject approaches (e.g., how does prefrontal cortex activity change in a person across experimental conditions?). Given its link to clinical variation, research that includes descriptions of cross-sectional brain-behavior relationships is often published in high-impact journals and improves chances of grant funding. As a result, researchers may feel systematic pressure to search for such relationships.

However, both the human brain and behavior are complex. The likelihood of finding a strong, direct mapping between any given brain region and behavioral measure across the general population is small. For example, depression has numerous symptoms, which can appear in different combinations, and are each likely driven by many brain areas and their intertwined functions. Perhaps it is unsurprising that the degree to which a complex behavioral measure tracks a single brain measure across people is typically small.

A new paper by Marek et al. (2022) quantifies this issue, demonstrating that many cross-sectional brain-behavior correlations in the general population are far smaller than previously acknowledged. The authors correlated brain and behavioral measures across people from three large consortia datasets (the Adolescent Brain Cognitive Development [ABCD] study, the Human Connectome Project [HCP], and the UK Biobank). Brain measures included structural MRI, resting-state fMRI, and (in supplemental analyses) task fMRI activations. Behavioral data included demographic, cognitive, and mental health/personality measures collected outside of the scanner to assess trait-like characteristics of each person. These brain and behavioral measures are common indices used across many papers in human neuroscience.

When examined in the ABCD ($N = 3{,}928$), brain-behavior associations were very small, topping out at $|r| = 0.16$ (mean $|r| = 0.01$, first percentile $|r| = 0.06$). These findings were irrespective of multiple comparisons correction, suggesting that targeted, hypothesis-driven searches with these measures are unlikely to find bigger effects. A similar distribution of effect sizes was seen in the HCP ($N = 1{,}200$) and UK Biobank ($N = 33{,}735$). Given these effect sizes, standard power analyses suggest that very large samples ($N > 1{,}000$) are needed to reliably correlate brain measures and behavior across the general population. Thanks to sampling variability, Marek et al. (2022) demonstrate that smaller samples produce a wide variety of (often incorrect) results, including apparently large-magnitude but non-replicable relationships.

Although a primary focus of this paper was on brain measures from resting-state fMRI, the findings do not indicate that resting-state fMRI is uniquely or even particularly poor in this regard. Indeed, effect sizes were generally larger for functional than structural MRI, larger for cognitive measures relative to measures of psychopathology, and larger for multivariate relative to univariate approaches. However, even the strongest multivariate effects found between resting-state fMRI and cognition required sample sizes of several hundred—more MRI data than many single investigators can feasibly afford to collect. Of course, even this massive investigation could not examine all possible measures employed in human neuroscience. It is possible that unexamined brain measures (e.g., from EEG recordings) and/or behavioral measures (e.g., from other cognitive tasks) could demonstrate superior effect sizes. However, Marek et al. (2022) found such ubiquitously small effects across such a broad range of measures that, in our view, we should assume that novel, unexamined cross-sectional associations will also be small

 © 2022 Elsevier Inc.

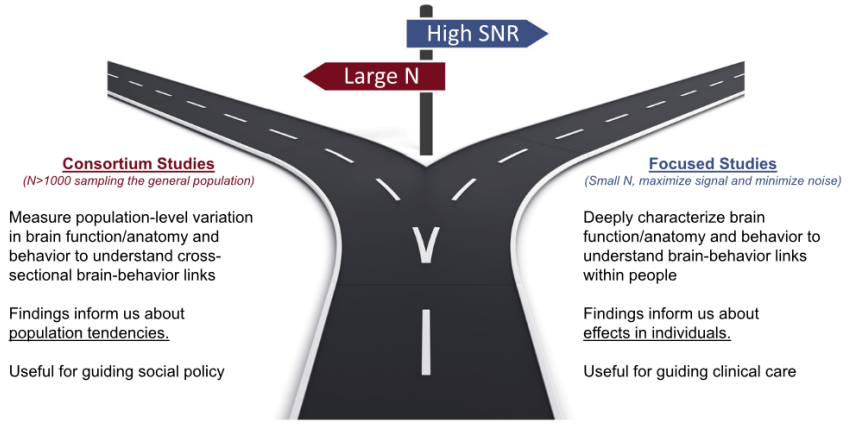

Figure 1. Two paths forward for reliable human neuroscience
Human neuroscience can move forward by embracing alternate study designs to identify replicable brain-behavior associations. We can sample broad trait-level effects in very large numbers of subjects (left path) to uncover low-magnitude but robust population-level associations that will inform social policy. Alternately, we can conduct small, focused designs that maximize signal and minimize noise through controlled, theoretically driven, repeated measures that deeply characterize individuals (right path) to describe larger-magnitude changes in brain and behavior induced by experimental interventions, development, or injury, which will inform clinical care.

unless well-powered and replicable evidence to the contrary is presented.

Alarmingly, these results, combined with publishing and funding incentives to include data with significant brain-behavior relationships, suggests that the literature (where typical sizes range from N = ~25–100) is likely enriched in underpowered, questionable, cross-sectional associations. Notably, this is true without presuming ill intent by researchers. Even when research groups eschew problematic data analysis practices (e.g., p-hacking, circular statistics, etc.), the simple mathematics of sampling variability prevent low-N cross-sectional brain-behavior correlations in the general population from replicating.

Continuing along the path of conducting underpowered cross-sectional research using similar measures to Marek et al. (2022) appears to be a dead end. In our view, the implications of this work leave us with two contrasting paths forward in finding reliable brain-behavior links (Figure 1): one based on consortia and the other based on smaller focused designs that maximize signal and minimize noise.

If we wish to continue using cross-sectional approaches as a primary means of understanding associations between brain and behavior, the findings of Marek et al. (2022) suggest that we should join forces across research groups to do consortium-level work. This approach, led by endeavors like the UK Biobank and ABCD, mimics the model of genomics or astrophysics, where large amounts of data are collected through joint endeavors. Single labs can then analyze public data from these large consortia. This approach is likely our best option for identifying broad cross-sectional brain-behavior associations generalizable to the population.

The Marek paper suggests that the effect sizes we find with well-powered consortium-level approaches will likely be small. This may be disheartening, and some may ask why we should care about (or fund efforts to discover) effects that explain so little variance in individual behavior. However, these effect sizes are still generally larger than those observed with well-established genome-wide association study approaches (e.g., Sniekers et al., 2017). Further, even small effects can have substantial impact at the population level. The strongest effects observed by Marek et al. (2022) are of a similar magnitude to associations between IQ and childhood lead exposure (Searle et al., 2014)—an effect that has driven substantial changes in public health policy. A consortium-level study that identifies effects of, for example, socio-economic status, stimulant use, or physical activity on brain function could have a similarly critical impact on policy decisions, even if the observed effect sizes are small.

However, consortium-level work also presents challenges and inherent limitations to the types of scientific questions that can be asked. As large consortia studying thousands of participants cost hundreds of millions of dollars to execute, consortia can be too risk averse to address novel experimental questions. A consortium will shy away from scientific approaches that have not been rigorously validated to avoid the risk of costly mistakes. Instead, emphasis is likely to be placed on common and well-established behavioral and brain measures, leaving entirely new questions without a path forward. Measures that require additional training for study personnel may also be de-emphasized—a trend seen in the preponderance of structural and resting-state MRI data collected in extant consortia compared to task fMRI.

Beyond scientific limitations, a shift toward large consortia raises issues regarding the distribution and equity of funding across the field. Consortium funding is typically awarded to established senior investigators. If funding priorities are shifted to consortium studies, it could impair the ability of early-stage investigators to obtain funding that allows them to establish independent laboratories, progress their careers, and achieve promotion. In the extreme case, a shift in funding priorities could squeeze out many junior investigators or leave them operating without independence within a consortium headed by a senior colleague. This could fundamentally change the career dynamics and organization of our field.

These possibilities are problematic. Neuroscience must be able to explore new scientific approaches and ask new questions, and early-stage investigators must be able to advance their careers with their own independent work, or our field will stagnate. Future funding cannot, therefore, be concentrated solely or even primarily on large consortia studies. Neuroscience must jointly pursue another, alternate path forward that asks new questions with high-powered, replicable methods.

In our view, a second viable path forward is to conduct smaller, focused studies that maximize effect sizes. For many readers, the take-home message of Marek et al. (2022) is about needed sample sizes. However, this work is really a story about statistical power, which is too low to identify cross-sectional associations between brain and behavior in the general population without thousands of participants. This doesn't mean that small samples should never be used or that prior small-sample literature may be casually dismissed. Rather, it means that small samples require higher-powered effects. To ask questions that cannot be studied in large consortia, neuroscience can embrace study designs that maximize statistical power by both revisiting methods used historically in cognitive neuroscience and advancing new individually focused approaches.

There are two primary ways that power can be increased via study design. The first is to increase the magnitude of the effects being studied (in signal-to-noise or "SNR" terms, to maximize the "signal"). To establish clinical utility, a study may test for cross-sectional associations between, for example, prefrontal cortex function and depressive tendencies among undergraduates, who are commonly used as a sample of convenience. But this approach is likely to minimize the size of cross-sectional associations because differences among healthy, nonclinical controls with uniformly high-normal intelligence are not expected to be especially large for either brain function or behavioral measures.

Effect sizes for brain-behavior associations are likely to be much larger when there exists substantial variation in both brain function and behavioral performance. Alternative study designs in this domain can maximize differences in the measures being tested through, for example, subject selection, experimental/causal manipulations, or a focus on theoretically motivated measures.

Between-group designs can focus on studying populations known to have large alterations in brain function, such as those with acquired brain injury or neurodegenerative disease. Indeed, the field of cognitive neuroscience began with studies of very large behavioral effects in lesion patients such as H.M. (Corkin, 1984), which established brain-behavior relationships with N = 1.

Increased signal can also be accomplished through within-subject designs by studying circumstances known to have large effects on behavior or cognition—or, even better, inducing large effects via experimental manipulation—and identifying brain measures associated with those changes. Strong theoretical frameworks can boost signals further by improving measurements and guiding interventions. Much of past task-based fMRI employs this powerful approach, resulting in foundational, well-replicated knowledge about many aspects of brain function. Examples include the organization of visual areas (Engel et al., 1997) and category selectivity in the brain (Kanwisher et al., 1997)—notable instances sufficiently powered to identify reliable effects in sample sizes of N < 10.

The second way to increase power is to minimize variation in brain or behavior measures that are unrelated to the association of interest (the "noise"). Some sources of noise (e.g., head motion; Ciric et al., 2017) are well known in neuroimaging research. Other sources of noise are less well appreciated. First, the fMRI signal itself is inherently autocorrelated and noisy. As a result, fMRI-derived measures are not reliable within subject unless extended/repeated data acquisition strategies are employed (Gordon et al., 2017), although advances in MRI technology may improve this issue (Lynch et al., 2020). Simulations from Marek et al. (2022) suggest that maximizing MRI and behavioral reliability alone could approximately double effect sizes.

Second, cross-subject heterogeneity confounds many behavioral and neuroimaging measures but can be ameliorated through controlled designs. For example, individuals differ in basic reaction time, so complex reaction time measures are compared to a baseline. A similar logic is applied in many task fMRI designs. However, other measures may not have an obvious within-subject "control" condition. MRI-derived measures also suffer from uncontrolled variation across subjects due to spatial variability of human functional neuroanatomy (Gordon et al., 2017). The most effective way to control for these various forms of cross-subject heterogeneity is to conduct brain-behavior associations using repeated measures within-subject designs and, at the extreme, focus analyses on individual-level "precision" data (Gordon et al., 2017). This person-centered approach controls for many potential sources of unwanted variation across individuals, including variation in genetics, experience, brain organization, and basic behavioral or physiological properties.

With this SNR-focused framework, we can consider the type of study best-suited for understanding brain-behavior associations within smaller samples. Ideally, such a study would examine brain signals associated with large alterations in behavior or cognition. It would use extended and/or longitudinal sampling of both neuroimaging measures and behavior. It would employ appropriate controls for motion and other sources of heterogeneity, and primary hypothesis testing would be conducted with repeated measures comparing against baselines, ideally even entirely within each individual. Not every research question can be developed into a study design that incorporates all of these features. However, the more that are incorporated, the larger study power one might expect. As an example, consider Newbold et al. (2020), in which a within-subject, repeated-scanning (>40 sessions/subject) design conducted with N = 3 examined brain measures before, during, and after a powerful behavioral intervention (temporary arm immobilization). With so many factors intrinsic to the study design contributing to increase signals and reduce noise, effect sizes of the behavioral intervention on brain function were massive, ranging from d = 1.0 to 3.8.

A small, focused study design that maximizes signal and minimizes noise is not inferior to a large consortium study. It is an alternate approach with equally high utility but for addressing different questions. While a cross-sectional consortium study can inform social policy, an SNR-focused study, by nature, focuses on individuals by asking questions about the presence, change, and manipulation of effects within a person. Such characteristics link SNR-focused designs closely to clinical care translation, which is inherently centered on individual patients. Consider a hypothetical consortium study that identifies significant

links between dorsolateral prefrontal cortex function and depressive symptoms in the general population, with effect sizes of $r \leq 0.16$. Such findings, while important, would be far from enabling improved diagnosis or prognosis for individual patients. In contrast, consider a hypothetical within-subject design that tracks how weekly measures of dorsolateral prefrontal function predict depressive mood before, during, and after transcranial magnetic stimulation treatment. Success of this study would have immediate implications for how we determine targets for intervention and how we understand whether a treatment is working for a given patient.

The findings of Marek et al. (2022) represent a fork in the road for neuroscience studies seeking to associate brain with behavior. It suggests we should focus our efforts on one of two extremes of experimental design, as appropriate for a given question—either large N to describe broad population-level brain-behavior associations or focused, lower-N, high-SNR designs to describe brain-behavior associations within individuals. Studies that walk an intermediate road and address both types of questions at once are likely not well adapted to either. Funding agencies should be wary of funding cross-sectional brain-behavior associations with sample sizes less than several hundred (without convincing evidence that an unusually large effect size will be achieved). Journals should be wary of manuscripts with cross-sectional brain-behavior correlations in smaller samples unless out-of-sample replication is achieved. No longer should such associations, presented without replication or consideration to power, be the gold standard for our field. Instead, greater value should be placed on powerful findings from high-signal, low-noise, within-subject designs that have consistently proven replicable in the field and allow for flexible testing of new findings and methods. In the end, multiple different approaches—based both on small, focused samples and large consortia—are necessary to understand the full range of how the brain is related to behavior.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020.

Corkin, S. (1984). Lasting consequences of bilateral medial temporal lobectomy: Clinical course and experimental findings in H.M. Semin. Neurol. *4*, 249–259. https://doi.org/10.1055/s-2008-1041556.

Engel, S.A., Glover, G.H., and Wandell, B.A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex *7*, 181–192. https://doi.org/10.1093/cercor/7.2.181.

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., et al. (2017). Precision functional mapping of individual human brains. Neuron *95*, 791–807.e7. https://doi.org/10.1016/j.neuron.2017.07.011.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997.

Lynch, C.J., Power, J.D., Scult, M.A., Dubin, M., Gunning, F.M., and Liston, C. (2020). Rapid precision functional mapping of individuals using multi-echo fMRI. Cell Rep. *33*, 108540. https://doi.org/10.1016/j.celrep.2020.108540.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. Nature *603*, 654–660. https://doi.org/10.1038/s41586-022-04492-9.

Newbold, D.J., Laumann, T.O., Hoyt, C.R., Hampton, J.M., Montez, D.F., Raut, R.V., Ortega, M., Mitra, A., Nielsen, A.N., Miller, D.B., et al. (2020). Plasticity and spontaneous activity pulses in disused human brain circuits. Neuron *107*, 580–589.e6. https://doi.org/10.1016/j.neuron.2020.05.007.

Searle, A.K., Baghurst, P.A., van Hooff, M., Sawyer, M.G., Sim, M.R., Galletly, C., Clark, L.S., and McFarlane, A.C. (2014). Tracing the long-term legacy of childhood lead exposure: A review of three decades of the port Pirie cohort study. Neurotoxicology *43*, 46–56. https://doi.org/10.1016/j.neuro.2014.04.004.

Sniekers, S., Stringer, S., Watanabe, K., Jansen, P.R., Coleman, J.R.I., Krapohl, E., Taskesen, E., Hammerschlag, A.R., Okbay, A., Zabaneh, D., et al. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nat. Genet. *49*, 1107–1112. https://doi.org/10.1038/ng.3869.