An Open Publishing Platform of
the Organization for Human Brain Mapping
APERTURE NEURO

COMMENTARY

Lost in transformation: fMRI power is diminished by unknown variability in methods and people

Peter A. Bandettini

Section on Functional Imaging Methods and Functional MRI Facility
National Institute of Mental Health, Bethesda, MD 20817, USA

Over the past 40 years, MRI has had immense clinical impact on patient assessment and care, as lesions associated with trauma or disease are readily identifiable in individual patients. Such an impact is an aspiration of many performing functional MRI (fMRI) or measuring subtle anatomic features as cortical thickness – toward assessment of psychiatric and/or developmental disorders. Associating MRI measures, including fMRI, with behavioral phenotypes across normal and pathological ranges, termed brain-wide association study (BWAS) has been deeply challenging due to the relatively small correlations coupled with substantial variation among individuals. The recent paper by Marek et al. (1) has shined a spotlight on this challenge through careful statistical analysis of cortical thickness and mostly resting state fMRI measures as they relate to population traits of psychosis and intelligence.

The conclusion, as stated in the title, is that due to the extremely small correlation values measured, the number of subjects required for reproducible BWAS is in the thousands. This title, while accurate in this context, unfortunately, catalyzed the popular media and even many fMRI practitioners to mistakenly paint a dismal picture of past results and future prospects for fMRI utility. The authors of the paper were careful to add clarifying paragraphs on the sustained importance of small-sample neuroimaging and possible ways to decrease variability; however, more should be said on these two topics.

The authors' primary message was to quantify and reinforce the need for larger sample sizes. Before we all start accumulating bucket loads of fMRI data in earnest or perhaps throw up our hands in despair, I believe we should focus on more finely and carefully dissecting and correcting sources of variability in acquisition, spatial normalization and registration, parcellation, paradigm design, population phenotyping, as well as subject performance in the scanner. We should also perhaps use the message of this paper to redirect some effort to areas of research where fMRI remains an established, powerful, and sensitive tool. These include longitudinal studies of brain plasticity with learning, experience, or recovery, the detailed mapping of functional hierarchy, organization and interaction among regions and layers, and the use of fMRI toward neurofeedback.

Regarding BWAS, quite a bit can and should be done that may improve results and decrease the numbers needed for reproducibility. Through better understanding the sources of variability, tractable insights on how individuals differ would likely be derived. Deep analysis of these sources of variability is beyond the scope of this paper; however, I would like to touch on a few.

Questions and perhaps some cognitive dissonance arose in many as they read this paper. Some may ask something along the lines of: "I see strong and clear activation in a single time series – even a single event – how can this signal require 1000's of subjects to get a result." This question speaks to the well-characterized strength and robustness of the fMRI signal change and "central-tendency" mapping and hints at a suspicion that something is wrong in the process of pooling, averaging, and comparing subjects. On the other hand, the reality may simply be that while the field can improve on the result shown, through better normalization or time series cleanup, the neuronal correlates (as measured with structural and functional MRI) of such complex behavioral measures may indeed naturally vary across subjects in such a way that we only capture the most prominent, and only with 1000's of subjects. This is importantly a statement about brain–behavior variability across humans and not about the sensitivity of fMRI.

The authors mentioned various methods for improving the quality of the data that include acquisition strategies such as multi-echo, denoising strategies, real-time quality control, and better phenotyping. All these approaches may indeed help. In my opinion, multi-echo will be particularly helpful, as would more deep and specific

phenotyping. Large groupings of what are likely highly inhomogeneous trait spectra (intelligence and psychosis) are used. A challenge in the field today is the reliance of likely outdated and imprecise phenotypical assessment. In the field of big data, the discussion has arisen of letting the data sort itself out based on what the strongest differences are and then using these results to refine phenotypical measures.

Another source of variance is the likely structural and functional spatial variability between individuals and the inadequacy of standard pipelines for controlling for this variation (2). The step of normalizing all brains to a standardized template and then parcellating the functional units for correlation analyses likely results in substantial mixing across parcels containing the fMRI resting state signal. The parcel sizes and locations generally match "functional units" in the brain. Larger parcels may partial-volume average diverse functional activity and connectivity within a parcel and result in more within-parcel but less across-parcel mixing-related diminished signal homogeneity across subjects, while smaller parcels may be more homogeneous in individuals but be completely washed out when averaging 1000's of individuals – due to inaccurate spatial normalization accounting for subject differences in anatomy. The problem can be further illustrated by the extreme example of say, attempting to use current spatial normalization methods to assess fMRI activation associated with individual digits or ocular dominance columns when looking at fMRI data between 1000's of subjects with amblyopia versus healthy controls. In both cases, all the meaningful signals would be completely averaged away by the natural variation in structure and function at this fine spatial scale. We should likely not assume that the most meaningful and strong BWAS need to occur at a low spatial resolution and have no significant spatial variation across individuals. This problem is further exacerbated by the well-known observation that networks and even parcellations dynamically reconfigure over time. Such reconfigurations may, in themselves, be useful measures to differentiate individuals, but in this processing pipeline, are lost.

The problems of imprecise spatial normalization and cross-subject registration are potentially addressable by further study of structural and functional spatial variation across individuals, and development of precise, likely nonlinear transformations that are unique to each individual. Approaches such as hyperalignment (3) also show promise for such studies. This problem also reveals an important difference between the concept of Gene-wide association studies (GWAS) and BWAS. An analogous situation would be if the essentially digital Genome-wide data had their single nucleotide polymorphisms shuffled with hundreds of adjacent polymorphisms over hundreds of places for each individual before any comparisons were made. One can imagine that in such a case, no inferences would be possible. Another point of power loss is in the use of resting state fMRI. It has clear utility because of its ease of implementation but is likely suboptimal for comparing across subjects. One can use the analogous concept of a cardiac "stress test" where only when stressed are pathologies revealed. Naturalistic stimuli in fMRI that are developed to draw out differences in individuals are being developed and have been shown to be more powerful than resting state studies (4).

Since the beginning of fMRI, the signal has been shown to be profoundly robust – so much so that, with a single event, the fMRI signal change was easily visible to the eye. Significant and reproducible maps of activation are able to be created with single scan sessions or even runs. When comparing sets of data across individuals, it is not the integrity of the fMRI signal itself that is the source of variability, it is likely the individual variations in behavior, structure, and functional location that wash out and distort the data. In addition, well-addressable factors such as physiologic noise, motion, and other artifacts contribute to a degree. Each individual's performance may vary throughout a scanning session, causing unknown variance that reduces sensitivity to differences. If the resting state is used, the resting state network configurations have long been shown to vary over time across at least 10 to 20 "states" that all may have different dwell times and duty cycles. Such individual variation in resting state dynamics has not been fully characterized. Averaging of resting state data across an entire run removes this potentially useful information and may skew results depending on state dwell time variation not associated with the population being compared. Finding differences is a difficult challenge since how precisely they manifest in time and space in the brain is still largely not understood. The Marek et al. study uses data that capture a sliver of this highly multidimensional space to explore. The spatial variation in structure and function in individuals, as it is observed in this study, may skew or wash out the meaningful differences, depending on the unique temporal reconfigurations of resting state signal, spatial normalization pipeline, and number of parcels used. Functional MRI and MRI only capture a narrow temporal and spatial scale, which may indeed reside outside where the more obvious differences are manifest.

To restate, fMRI can be used to map detailed structure/function relationships and can be used to track changes in learning and adaptation over time scales of seconds to years in individual subjects; however, group comparisons are much more challenging for three primary reasons: The first is that spatial normalization across subjects remains imprecise. The second is that the most informative aspect of the fMRI signal may be overlooked. The third is that the true spatial variation of fMRI measures across subjects with similar traits may indeed be substantial – defying most attempts at pooling, averaging, and differentiating.

While Marek et al. summarize the problem as it exists with the current state of the art of phenotypic delineation, paradigm design, acquisition, denoising, and spatial normalization, it is only a snapshot of a rapidly evolving set of approaches. For over 30 years, the fMRI signal has been repeatedly shown to be highly robust and sensitive to neuronal activity. Comparison of groups and individuals requires more effort to more fully characterize and adjust for group and individual variation in behavior, brain structure, and functional localization.

REFERENCES

1. Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., et al. Reproducible brain-wide association studies require thousands of individuals. Nature 2022;603(7902):654–60.
2. Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., et al. Moving beyond processing and analysis-related variation in neuroscience. bioRxiv; 2021 [cited 2022 Apr 22]. p. 2021.12.01.470790. Available from: https://www.biorxiv.org/content/10.1101/2021.12.01.470790v1.
3. Haxby, J.V., Guntupalli, J.S., Nastase, S.A., Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. eLife 2020. Available from: https://elifesciences.org/articles/56601.
4. Finn, E.S. Is it time to put rest to rest? Trends Cogn Sci 2021;25(12):1021–32.

