A Reckoning and Research Agenda for Neuroimaging in Psychiatry

Amit Etkin, M.D., Ph.D.

Human neuroimaging has been a core component of both research in psychiatry and conceptual models of the brain circuit-level mechanisms underlying psychopathology. Despite landmark neuroimaging research over the past 25 years, we still lack the level of precision and insight needed for bringing neuroimaging tools into clinical care contexts. This brief review examines historical research trends in psychiatric neuroimaging, as well as the basic assumptions underlying current efforts, in order to understand factors that have limited the impact of neuroimaging efforts thus far. These factors include the pitfalls of case-control designs, confounders inherent in associational research approaches, and the challenges in embracing fully data-driven analyses. Several critical gaps emerge, the addressing of which could provide the critical new insights that have long been sought from neuroimaging. These include transitioning from group-to individual-level analyses (and through this to intervention studies carried out robustly at the level of individual patients), building "big data" from a longitudinal perspective and not only a cross-sectional one, a greater focus on identifying causal mechanisms, and the development of tools such as electroencephalography in addition to the dominant MRI methods to aid translation to real-world clinical care. Despite the still-unrealized potential of psychiatric neuroimaging, there is now much to be excited about as previous learnings are converted into fundamentally new directions. Indeed, we may now be at an inflection point for neuroimaging if the typical study designs are left in the past and the field systematically and thoughtfully embraces the challenges that the past 25 years of research have now made apparent.

Am J Psychiatry 2019; 176:507–511; doi: 10.1176/appi.ajp.2019.19050521

Neuroimaging has long been the primary tool for understanding the biological basis of psychopathology. Using that tool, a generation of researchers has sought the neural correlates of psychiatric diagnoses, risk states, and treatment effects. The cost of doing this type of work has come down as technology has improved and become more accessible, leading to progressively larger or more diverse clinical samples being studied. Moreover, organizing initiatives such as the Research Domain Criteria (RDoC) project from the National Institute of Mental Health have strongly influenced the conceptual models driving neuroimaging research (1). Despite these efforts, we are still far from a neuroimaging-based diagnostic, predictive, or surrogate-endpoint marker. Now is therefore a good point for reflecting on what we have learned from these neuroimaging efforts, assessing where collective misjudgments or missed opportunities have occurred, and examining how the field can consolidate these insights to make more fundamental progress in the coming years. This review aims to address these questions in a general manner, acknowledging that such generalities will also necessarily miss specific findings that may run counter to the conclusions drawn. Nonetheless, the thesis of this brief review is that by examining the overall state of the science in neuroimaging, as exemplified as well by articles in this issue of the *Journal*, important elements of a future-looking research agenda can emerge.

FROM SMALL-N CASE-CONTROL STUDIES TO LARGE TRANSDIAGNOSTIC SAMPLES

For many of its early years, in large part because of the cost and burden of positron emission tomography studies, neuroimaging samples were very small, and in retrospect, unrealistically powered to detect true effect sizes. This trend continued for many years afterward as functional MRI (fMRI) and structural MRI took over as the preferred neuroimaging tool. Sample sizes have increased from ~20 per group to now routinely hundreds of individuals, which in turn has improved statistical detection power, as illustrated by reports in this issue from Wannan et al. (2) and Lizano et al. (3).

Nonetheless, the dominant model for psychiatric neuroimaging has largely remained the case-control study. Specifically, patients are enrolled based on a specific clinical diagnosis and are compared with healthy individuals, but typically not to another clinical group. Imaging is usually also performed at a single time point. Inherently, therefore, such studies assume that a clinical group, defined on the basis of

specific clinical criteria, will define a mechanistically meaningful study sample that will differ from healthy individuals based solely on the criteria along which they were selected. That is, by enrolling people meeting criteria for a particular diagnosis, the case-control contrast captures the essence of "depression" or "schizophrenia" as a clinical construct. Moreover, while a small number of more recent studies, influenced by frameworks such as RDoC, have examined clinical cohorts using continuous measures of symptoms (rather than dichotomizing between patients and controls), these nonetheless represent a minority. Such studies also still assume that symptoms are the best organizing principle for neuroimaging investigations. What, then, have case-control studies taught us about psychiatrically relevant neural dysfunction?

Large-scale meta-analysis of both brain structure and brain function, whether assessed at rest or through cognitive or emotional tasks, have found very few differences between seemingly phenotypically distinct diagnoses (4–6). Rather, the repeated finding across these modalities has been one of a common circuit dysfunction across psychiatric diagnoses, despite the fact that no single symptom is shared across the diagnostic criteria for these conditions (although all share the fact that they impair daily functioning in some way). In fact, separate meta-analyses of brain structure and cognitive task activation have shown a direct overlap between regions showing disorder-general dysfunction (4). Studies of large samples with diverse clinical diagnoses have found similar nonspecificity (7).

It is not just neuroimaging that has failed to find robust correlates of specific diagnoses. Genetic studies have repeatedly found very similar results, wherein genetic risk is largely shared among seemingly divergent clinical conditions (8). While it may be that both neuroimaging and genetics are only sensitive to disorder-general dysfunction, a more plausible explanation is that the clinical definition of psychiatric disease has failed to identify individuals with neural dysfunction attributable to specific clinical states. Indeed, it is this perspective that led to the development of RDoC and related frameworks. However, we still largely lack evidence that symptom-defined clinical groups, when examined dimensionally instead of categorically, necessarily have greater external validity when examined through the lens of neuroimaging. In other words, the dimensional focus of RDoC, as a point of contrast with the categorical definitions of DSM, remains a hypothesis in need of testing rather than already being a solution to the failures of the case-control study design in revealing the neural basis of different clinical presentations.

What, then, have we learned from the past two decades of case-control psychiatric neuroimaging? One view might be that we have learned a lot less than the work we have put in might presume we should have, as no consistent diagnosis-related neural signature has emerged. An alternative view, however, would acknowledge this fact but also argue that in arriving at this point, we are now much better positioned to know which questions we should be asking in neuroimaging studies, some of which are detailed below. Nonetheless, given this inflection point in the field, we must also seriously ask ourselves, Are we ready to retire the psychiatric case-control study using DSM-based diagnostic definitions?

THE EMERGENCE OF LARGE SAMPLES AND DATA-DRIVEN HUMAN NEUROSCIENCE

With the greater ease of use and access to neuroimaging technologies, individual studies have rapidly grown in size, often to the low hundreds of participants. At the same time, interest in data sharing has increased, spurred also by changes in data access rules by major funders. Simultaneously, mega-studies such as the Human Connectome Project, the UK Biobank, and the Adolescent Brain Cognitive Development studies were created to provide high-quality, publicly accessible data sets with data availability as close to collection as practically possible. At least for some questions, data size is no longer an issue, suggesting that data sharing should be highly encouraged across the field. The Hoogman et al. report in this issue of the *Journal* (9) illustrates this shift, with the authors' primary analysis involving over 4,000 individuals from 36 imaging centers. However, by having large samples for analysis now, attention in such studies should shift to a primary focus on effect size (and clinical significance) rather than statistical significance. As is typically the case in other large-scale structural comparison studies, the largest effect size (d) observed in the primary analysis of Hoogman et al. was 0.21, which is below even the 0.3 threshold for what is considered a "small" effect size (with a d value of 0.5 reflecting a clinically significant "medium" effect size).

With both data sharing and mega-studies, however, come several inherent limitations that must be considered. First, by necessity, pooling data from different studies requires using neuroimaging protocols that are common across the contributing data sets. In practice, this tends to mean either structural MRI or resting-state fMRI (i.e., fMRI data collected while participants lie quietly in the scanner without doing a specific task). Also, large samples tend to be substantially skewed toward healthy individuals rather than treatment-seeking patients, let alone those with severe mental illnesses. With advances (and broad interest) in machine learning methodologies, such large data sets are nonetheless an excellent substrate for more purely data-driven approaches to identifying brain-based "biotypes." Several recent studies have demonstrated interesting insights possible by leveraging such data sets. Moreover, it is already clear in these early efforts that the DSM-defined categories used in previous case-control studies comprised multiple biologically distinct neural phenotypes (10–12). Such biological heterogeneity appears to be the rule rather than the exception, and it helps illustrate at least part of what case-control studies in the past have missed.

For data-driven definitions of biotypes to succeed, however, several key issues must be closely attended to and systematically addressed. If biotypes (or continuous dimensions of neural functioning, rather than categorical classes) are to be of use in psychiatry, they must have some external validity and utility. That is, something clinically and/or mechanistically meaningful must differ as a function of such biotypes. A common challenge is that clinical symptoms may not differ between biotypes, with clinical symptomatology remaining a frequent default validation perspective in psychiatry. Indeed, there is a certain circularity in using clinical symptoms to validate data-driven biotypes if that is intended as a contrast to symptom-driven analyses. Instead, it may be that behavioral task performance or some other objective biological measure may differ between biotypes. Alternatively, clinical differences may exist but only become evident if the response to treatment is examined (i.e., different treatment outcomes across biotypes). Unfortunately, large-N studies typically lack information about treatment outcomes in a systematic manner. Similarly, the significance of biotyping efforts that pool across both healthy and clinical groups can be unclear. That is, while the biotyping goal of identifying more biologically homogeneous groups of individuals may be achieved, this may come at the cost of clinical significance if biotypes comprise both healthy and psychiatrically ill individuals. Hence, even assuming analytic robustness and replicability of a particular set of biotypes, how such data-driven approaches can be linked back to clinically meaningful matters such as treatment remains a critical and open question. This issue is illustrated by Hawco et al. in this issue of the *Journal* (13).

Given these challenges, a purely data-driven analysis of large data sets, even those with sufficient representation of clinical populations, may find it hard to converge on a set of consistent and clinically meaningful answers. Biotypes may furthermore differ depending on the data type and analytic approach. Such "battles of the biotypes" may lead to further fractionation of the literature, and with it greater difficulty advancing our understanding using neuroimaging. This is a particular concern if biotypes are defined in a manner that is hard for others to replicate and test in their samples.

One approach to advancing a data-driven perspective, but still realizing clinically meaningful goals, is to consider which measures could serve as the most relevant anchors for analysis. For example, is the goal to identify neural factors that distinguish groups of psychiatrically ill individuals from healthy individuals? Is the goal to inform treatment selection or therapeutic development? Given the uncertainty with which a data-driven biotype will be able to map onto these questions, designing collection efforts and data-driven analyses with more explicit anchoring on these clinical questions may be needed to yield the most fruitful answers. One area from which such learnings can be taken is the study of dementia, as illustrated by Licher et al. in this issue of the *Journal* (14).

NEW FRONTIERS FOR POTENTIAL BREAKTHROUGHS IN PSYCHIATRIC NEUROIMAGING

Having considered some of the lessons above in reflecting on the history of and current trends in psychiatric neuroimaging, several areas appear surprisingly understudied but carry substantial potential for achieving the breakthroughs long sought by neuroimaging. One conclusion in contrasting the failure of current group-level definitions (i.e., diagnosis based) to yield actionable insights, along with the broad enthusiasm around machine learning methodology, is that studies and analytic approaches may find greater success by focusing on the level of the individual. For this to succeed, a greater focus must be placed on establishing (and increasing) the retest reliability of neuroimaging measures. While the reliability of structural imaging is high, that of resting-state fMRI is low to moderate, and that of task-based neuroimaging can be highly variable.

A focus on the individual, however, has the advantage of opening up entirely new approaches in psychiatric studies, including facets that may reveal insights obscured by the nature of cross-sectional studies. For example, the current excitement over "big data" in practice presumes that the "big" dimension is one involving neuroimaging of many individuals, assessed at a single time point. Little attention has been paid to examining the orthogonal "big" dimension, namely, many neuroimaging time points for single individuals. To understand the profound implications of this shift, in which dimension is sampled most heavily, consider the following toy example (15). When examining the relationship between typing speed and errors, if a cross-sectional study is undertaken, then a negative relationship may well be observed—that is, those who type fastest make the fewer errors. By contrast, if a within-individual study is undertaken, a positive relationship may be found. This apparent fundamental divergence can be explained by the different sources of variation captured by these two study designs. In the cross-sectional study, the negative relationship reflects the fact that better typists can type faster and make fewer mistakes in doing so, while the individual-level study reflects the fact that as anyone is made to type faster, they will make more mistakes. Such failures in group-to-individual generalization have been noted as a particular challenge in research on humans (16).

Seminal work in the area of individual-level neuroimaging, using repeated assessments of a single highly sampled individual, has demonstrated relationships between session-to-session variations in self-report and physiological characteristics and neuroimaging signals (17). Other work has used individual-level repeated neuroimaging as a strategy to get highly stable signal estimates, and through this to demonstrate finer parcellations in network structure than previously appreciated (but not relating these to session-to-session variations in other features) (17). Nonetheless, we still lack any serious large-scale effort to do such "deep neural

phenotyping" in psychopathology. If successful, however, such an approach can help establish neuroimaging as a "brain vital" akin to typical vital signs routinely monitored in medical care. Likewise, longitudinal deep phenotyping efforts can support much-desired N-of-1 studies. In this type of study design, individuals may receive multiple acute interventions, along with control interventions, so that the effect of the intervention can be statistically tested for that individual. Because of their rigor and applicability to the individual, N-of-1 studies are considered to be at the top of the hierarchy of evidence-based medicine methods (19). To be successful, however, the intervention must be short-term and reversible or transient, and the outcome measure robust and sensitive. Although we have not reached this point with neuroimaging, a focus on individual-level imaging has the potential to be the path in psychiatry toward N-of-1 studies in a way that clinical constructs such as mood or psychosis have not delivered on.

Ultimately, while big data efforts, whether they are cross-sectional or longitudinal, may help build stronger associations using neuroimaging signals, they do not address a deeper problem with neuroimaging: the building of ever-stronger associations may not bring us any closer to understanding causal circuit-level mechanisms in psychopathology (20). In fact, there is ample evidence of strong associations failing to reveal causal factors, and potentially even proving to be misleading. One notable example in medicine is the relationship between hormone replacement therapy and coronary heart disease (21). Epidemiological studies suggested that hormone replacement therapy is associated with lower risk of heart disease, but when randomized controlled trials were performed, a surprising increased risk of heart disease was observed. In the context of psychiatric neuroimaging, therefore, the gap between the typical correlative study and the types of studies needed to demonstrate causality should raise substantial concern. There are many ways in which causality can be demonstrated, most saliently for neuroimaging by use of targeted mechanistic perturbations such as through transcranial magnetic stimulation (TMS) (either separate from or concurrent with neuroimaging) and pharmacology, to name just two tools. An example is seen in Brady et al. in this issue of the *Journal* (22). The great benefit of increased focus on dissection of causal relationships is that this more naturally yields a bridge between studies characterizing patients and opportunities for novel circuit-directed therapeutic interventions. Conceptually, the line between short-term interventions intending to reveal causal relationships in neuroimaging data and lasting therapeutic interventions with clinical impact may be more in degree than in kind. That is, treatment may be achieved by simply repeating the short-term intervention multiple times (e.g., one repetitive TMS [rTMS] session probing a circuit but a course of multiple rTMS sessions comprising treatment).

Finally, another striking facet of psychiatric neuroimaging research over the past two decades has been its relative disconnection from real-world clinical care. Neuroimaging studies have been carried out in the ideal environments of academic laboratories and research-dedicated scanners, often in unmedicated individuals who themselves may not be treatment seeking. As such, the applicability of current neuroimaging findings to real-world clinical care, and hospital-grade scanners, remains largely unknown. This gap, which threatens not only to restrict the practical utility of neuroimaging-based insights but also to produce potentially misleading conclusions, relates to a number of factors. In part because of MRI's greater spatial resolution, along with historical trends in tool use, MRI-based studies have attracted the vast majority of funding and attention in the field. Beyond the cost and nonportability of MRI, however, many fundamental questions remain about the impact of variations in data acquisition platforms and methods, as well as data analytic methods, which must all be solved in order to deploy this tool in clinical practice—much in the same way as required for any clinical test. Alternative approaches, such as EEG, which have been the focus of more intense study in the past, may prove pivotal in connecting neuroimaging research to biomarkers capable of large-scale implementation. Indeed, an influx of new methods for EEG (which measures electrical potentials in the brain) or magnetoencephalography (MEG; which measures the magnetic field created by electric potentials in the brain) holds great promise in allowing EEG to capture the types of network connectivity signals that are currently the primary purview of fMRI (23–25). Issues like platform dependence of EEG can be more readily dealt with given the lower price point for EEG than MRI equipment, and EEG has been routinely used to measure seizure activity for decades in neurology. Some EEG measures, such as mismatch negativity (an EEG signal deflection denoting that an unexpected event occurred), have also been extensively used and been found to be extremely reliable and very clinically relevant (26). The spatiotemporal richness of EEG may also prove to be a particular boon for machine learning methodologies, thus aligning this tool well with current trends in psychiatric data science.

CONCLUSIONS

The past two decades have seen neuroimaging come of age as a tool critical to the future of biological psychiatry. However, its impact remains limited in terms of producing new key insights, let alone clinically usable biomarkers. Through an honest reflection on trends and approaches the field has taken to date, and reckoning with their assumptions and impact, this review argues that we may be at an important inflection point for the field. By embracing what we have learned from past efforts and appraising critical gaps in the logic pervading current efforts, we can discern numerous directions that new research can take that may prove instrumental in achieving the neuroimaging-guided breakthroughs long sought by the field.

AUTHOR AND ARTICLE INFORMATION

The Department of Psychiatry and Behavioral Sciences and the Wu Tsai Neurosciences Institute, Stanford University, Stanford, Calif.; and the Veterans Affairs Palo Alto Healthcare System and the Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Palo Alto, Calif.

Send correspondence to Dr. Etkin (amitetkin@stanford.edu).

Dr. Etkin holds equity in Mindstrong Health and Akili Interactive.

Accepted May 20, 2019.

REFERENCES

1. Insel T, Cuthbert B, Garvey M, et al: Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 2010; 167:748–751
2. Wannan CMJ, Cropley VL, Chakravarty MM, et al: Evidence for network-based cortical thickness reductions in schizophrenia. Am J Psychiatry 2019; 176:552–563
3. Lizano P, Lutz O, Ling G, et al: Association of choroid plexus enlargement with cognitive, inflammatory, and structural phenotypes across the psychosis spectrum. Am J Psychiatry 2019; 176:564–572
4. Goodkind M, Eickhoff SB, Oathes DJ, et al: Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 2015; 72:305–315
5. McTeague LM, Huemer J, Carreon DM, et al: Identification of common neural circuit disruptions in cognitive control across psychiatric disorders. Am J Psychiatry 2017; 174:676–685
6. Sha Z, Wager TD, Mechelli A, et al: Common dysfunction of large-scale neurocognitive networks across psychiatric disorders. Biol Psychiatry 2019; 85:379–388
7. Baker JT, Dillon DG, Patrick LM, et al: Functional connectomics of affective and psychotic pathology. Proc Natl Acad Sci USA 2019; 116:9050–9059
8. Anttila V, Bulik-Sullivan B, Finucane HK, et al: Analysis of shared heritability in common disorders of the brain. Science 2018; 360(6395):eaap8757
9. Hoogman M, Muetzel R, Guimaraes JP, et al: Brain imaging of the cortex in ADHD: a coordinated analysis of large-scale clinical and population-based samples. Am J Psychiatry 2019; 176:531–542
10. Clementz BA, Sweeney JA, Hamm JP, et al: Identification of distinct psychosis biotypes using brain-based biomarkers. Am J Psychiatry 2016; 173:373–384
11. Drysdale AT, Grosenick L, Downar J, et al: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 2017; 23:28–38
12. Etkin A, Maron-Katz A, Wu W, et al: Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. Sci Transl Med 2019; 11(486):eaal3236
13. Hawco C, Buchanan RW, Calarco N, et al: Separable and replicable neural strategies during social brain function in people with and without severe mental illness. Am J Psychiatry 2019; 176:521–530
14. Licher S, Leening MJG, Yilmaz P, et al: Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. Am J Psychiatry 2019; 176:543–551
15. Hamaker EL: Why researchers should think "within-person": a paradigmatic rationale, in Handbook of Research Methods for Studying Daily Life. Edited by Mehl MR, Conner TS. New York, Guilford, 2012, pp 43–61
16. Fisher AJ, Medaglia JD, Jeronimus BF: Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci USA 2018; 115:E6106–E6115
17. Poldrack RA, Laumann TO, Koyejo O, et al: Long-term neural and physiological phenotyping of a single human. Nat Commun 2015; 6:8885
18. Braga RM, Buckner RL: Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. Neuron 2017; 95:457–471.e5
19. Guyatt GH, Haynes RB, Jaeschke RZ, et al: Users' Guides to the Medical Literature, XXV: evidence-based medicine: principles for applying the Users' Guides to patient care. JAMA 2000; 284:1290–1296
20. Etkin A: Addressing the causality gap in human psychiatric neuroscience. JAMA Psychiatry 2018; 75:3–4
21. Lawlor DA, Davey Smith G, Ebrahim S: Commentary: The hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol 2004; 33:464–467
22. Brady RO Jr, Gonsalvez I, Lee I, et al: Cerebellar-prefrontal network connectivity and negative symptoms in schizophrenia. Am J Psychiatry 2019; 176:512–520
23. Mantini D, Perrucci MG, Del Gratta C, et al: Electrophysiological signatures of resting state networks in the human brain. Proc Natl Acad Sci USA 2007; 104:13170–13175
24. Van de Ville D, Britz J, Michel CM: EEG microstate sequences in healthy humans at rest reveal scale-free dynamics. Proc Natl Acad Sci USA 2010; 107:18179–18184
25. Hipp JF, Hawellek DJ, Corbetta M, et al: Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat Neurosci 2012; 15:884–890
26. Naatanen R, Kujala T, Light G: The Mismatch Negativity: A Window to the Brain. Oxford, UK, Oxford University Press, 2019