June 7th, 2021

Michael B Eisen

Editor-in-Chief of eLife

HHMI, University of California, Berkeley

Dear Professor Eisen,

Please consider for publication our manuscript titled “Effect sizes and test-retest reliability of the fMRI-based Neurologic Pain Signature” in *eLife*. This paper is organized as a research article. fMRI research is at a critical juncture. Many large-scale studies have been afforded major funding in recent years. One of the foundational reasons for these substantial efforts is the promise to yield neuroimaging-based biomarkers related to human health and disease. Reliability has traditionally been a cornerstone of measurement theory when validating all types of measures in the behavioral and medical sciences. Though there has been longstanding interest in the reliability of fMRI measures, recent studies have brought the issue to the forefront. For example, Elliott et al. (2020) published a meta-analysis showing the task-related fMRI measures have poor reliability. Their work was construed as having broad applicability to fMRI measures in general and is a challenge that the field must address.

We address the topic of fMRI reliability in relation to fMRI-based biomarkers of health-related processes in this manuscript. In the largest analysis to date - on fMRI data from N = 364 participants in 10 studies - we analyze the test-retest reliability of the “Neurologic Pain Signature (NPS),” a measure related to evoked nociceptive pain (Wager et al. 2013, NEJM). The NPS is a particularly salient test case because this multivariate fMRI pattern was explicitly designed as a biomarker of a clinically relevant process suitable for clinical and translational use. And to our knowledge, it is the most widely validated neuroimaging-based biomarker to date (it has been tested in publications on >40 distinct cohorts).

We analyze the overall test-retest reliability and analyze several important variables that contribute to reliability, including effect sizes in predicting symptoms (i.e., pain) within- and between-person, the amount of data aggregation (i.e., the number of trials used to estimate a measure), and the time from test to retest. Overall, we find that test-retest reliability is excellent when sufficient data are collected, and the findings provide guidelines for future studies. Perhaps most interesting from a theoretical perspective, we find that while the brain measure (NPS) and pain reports were highly reliable and strongly linked in within-person analyses (with very large effect sizes), they were only modestly correlated in between-person analyses. This implies that brain and self-report measures of pain actually measure different constructs when it comes to inter-individual variability and are likely susceptible to different and complementary sources of bias and measurement error. This finding bears on the utility of neuroimaging-based measures (and other biomarkers) across mental health and related conditions, as brain imaging can provide a unique window into individual differences in pathophysiology that is not captured by self-report. We believe that our research will be of interest to the readership of *eLife*, as they connect with fMRI research, and more broadly to the public at large due to the recent debate of fMRI reliability (Kragel et al., 2021).

Thank you for your time and consideration.

Sincerely,

Tor Wager, Ph.D.

Director, Cognitive and Affective Neuroscience Laboratory

Diana L. Taylor Distinguished Professor

Department of Psychological and Brain Sciences

Dartmouth College