Published in final edited form as:
J Pain. 2014 October ; 15(10): 1008–1014. doi:10.1016/j.jpain.2014.06.011.

TEST-RETEST RELIABILITY OF PAIN-RELATED BRAIN ACTIVITY IN HEALTHY CONTROLS UNDERGOING EXPERIMENTAL THERMAL PAIN

Janelle E. Letzen, M.S.[1], **Landrew S. Sevel, B.A.**[1], **Charles W. Gay, D.C.**[2], **Andrew M. O'Shea, M.S.**[1], **Jason G. Craggs, Ph.D.**[1], **Donald D. Price, Ph.D.**[3], and **Michael E. Robinson, Ph.D.**[1]

[1]Department of Clinical and Health Psychology, University of Florida, Gainesville, FL

[2]Department of Rehabilitation Sciences, University of Florida, Gainesville, FL

[3]Department of Dentistry, University of Florida, Gainesville, FL

Abstract

Although functional magnetic resonance imaging (fMRI) has been proposed as a method to elucidate pain-related biomarkers, scant information exists related to psychometric properties of fMRI findings. This knowledge is essential for potential translation of this technology to clinical settings. The purpose of this study was to assess the test-retest reliability of pain-related brain activity and how it compares to the reliability of self-report. Twenty-two healthy controls (mean age = 22.6 years, SD = 2.9) underwent three runs of an fMRI paradigm that used thermal stimuli to elicit experimental pain. Functional MRI summary statistics related to brain activity during thermal stimulation periods were extracted from bilateral anterior cingulate cortices and anterior insula. Intraclass correlations (ICC) were conducted on these summary statistics and generally showed "good" test-retest reliability in all regions of interest (ICC range = .32 – .88; mean = .71); however, these results did not surpass ICC values from pain ratings, which fell within the "excellent" range (ICC range = .93–.96; mean = .94). Findings suggest that fMRI is a valuable tool for measuring pain mechanisms, but did not show an adequate level of test-retest reliability in this study to potentially act as a surrogate for individuals' self-report of pain.

Keywords

fMRI; Test-Retest Reliability; Anterior Cingulate Cortex; Insula

Corresponding Author: Michael E. Robinson, PhD, University of Florida, 101 South Newell Drive, Rm 3151, P.O. Box 100165, Gainesville, FL 32610-9165, Tel: 325-273-5220, Fax: 352-273-6156, merobin@ufl.edu.

Disclosure
There is no conflict of interest among authors.

Introduction

Recent papers have proposed functional magnetic resonance imaging (fMRI) as a method to elucidate an objective biomarker for the diagnosis of pain syndromes.[1,5,6,32,33] One justification made for using such a biomarker in clinical practice is that the self-report of pain is unreliable, which can make diagnosis and treatment difficult. However, studies have found high test-retest reliability of subjective pain ratings for both acute[4] and chronic[17] pain, whereas this information is lacking for pain neuroimaging.[28] The reliability of fMRI findings in the study of pain is essential to determine before potential translation of this technology to clinical practice, as reliability establishes the upper bound for validity.

Test-retest reliability is a measure of the extent to which scores are consistent and free from error.[19] It is important to note that an individual's pain experience is not static over time, and pain intensity or unpleasantness can fluctuate.[28] However, the degree to which these pain ratings vary is predictable, and therefore does not represent error.[30] Although the experience of pain can vacillate over time, subjective pain measures have been shown to reliably capture an individual's pain experience. Williamson and colleagues[34] conducted a meta-analysis of reliability studies on three commonly used rating scales: visual analogue scales (VASs), numerical rating scales, and verbal rating scales. The authors concluded that all three instruments had strong reliability across studies and were acceptable for clinical use. VASs showed the highest reliability across time points, with intraclass correlation (ICC) coefficients ranging from 0.97 to 0.99.[4,15] Jones and colleagues[18] found high repeatability of self-reported pressure pain thresholds across four consecutive days in pain-free women, with kappa coefficients ranging from 0.94–0.97. Commonly used questionnaires, such as the McGill Pain Questionnaire, have also been shown to have high reliability, with ICC coefficients ranging from 0.89 to 0.96 for total, sensory, affective, and average pain scores.[16] In general, higher reliability is found within shorter time spans due to the fluctuation of pain itself.[19]

Although the reliability of subjective pain report has been examined among different measures, there is a paucity of reliability studies in pain neuroimaging. In general, few studies have examined test-retest reliability of fMRI data, which could be lower than commonly expected in the field.[3] Brain activity within the default mode network demonstrated good reproducibility over three separate time points.[22] However, Brandt and colleagues[7] examined the test-retest reliability of a novelty encoding paradigm, and concluded that results were difficult to interpret at the single-subject level due to poor reliability. In a meta-analysis of fMRI reliability studies, Bennett and Miller[3] found an average ICC coefficient of 0.5 across several cognitive tasks in healthy controls. The authors also concluded that test-retest reliability was typically poorer among studies of clinical populations.

To the best of our knowledge, no studies have examined test-retest reliability of brain activity associated with pain processing, which is imperative to determine before this technology can be used clinically, as suggested by others.[5,6,20,33] The purpose of the present study was to measure the test-retest reliability of two brain regions associated with pain processing in healthy controls, and examine it compared to reliability of self-report. A meta-

analysis examining experimental pain fMRI studies showed that bilateral anterior insula (aINS) and right anterior cingulate cortex (ACC), among two other regions, had the highest likelihood of being activated by noxious stimuli.[13] We limited our a priori regions of interest (ROIs) to the ACC and aINS because they were suggested to best reflect pain perception[1] and sensitivity to changes in self-report,[8] respectively. While the ACC is involved in the affective component of pain processing, the aINS is involved in both affective and cognitive-evaluative components of pain.[13]

Methods

This study is a secondary data analysis from a larger, NIH-funded fMRI project examining mechanisms and temporal properties of placebo analgesia. For the parent study, individualized “pain” and “placebo” temperature thresholds were established during a screening visit using VAS responses to thermal quantitative sensory testing (QST) outside of the MRI scanner. Individuals who qualified to participate in the study then completed one baseline fMRI visit wherein only thermal “pain” temperatures were applied, with no placebo conditioning or other manipulation. This baseline visit was followed by two additional scanning sessions, each separated by one week, and participants underwent placebo conditioning prior to either the second or third scanning session. Before each scanning session, participants completed the State-Trait Anxiety Inventory (STAI) and the Pennebaker Inventory of Limbic Languidness (PILL). Data included in the present analyses were from the parent study’s baseline visit, and only represent brain activity and self-report associated with thermal, experimental pain. Methods described below represent procedures used for the baseline visit.

Participants

Data from 22 healthy, pain-free individuals were analyzed in this study (mean age = 22.6 y, SD = 2.9; 13 females). Nine participants identified as Caucasian, four as Asian, five as Hispanic, and four as African American. Participants were excluded if they met the following criteria: 1) current enrollment in another research study that could influence participation in the present study, 2) use of pain-related medications that could not be stopped seven days prior to testing (e.g., NSAIDs, antihistamines, antidepressants, anti-convulsants, migraine medications, and cough suppressants), 3) history of psychiatric, psychological, or neurologic disorder, as well as medical conditions associated with chronic pain, 4) current medical condition that could affect study participation, 5) positive pregnancy test result in females, 6) presence of ferromagnetic metal within the body, and 7) inability to provide informed consent. The parent study was approved by the University of Florida Institutional Review Board. All participants provided written informed consent.

Experimental Materials

Thermal stimuli during fMRI scanning periods were delivered using an MR-compatible, peltier-element-based stimulator (Medoc Thermal Sensory Analyzer, TSA-2001, Ramat Yishai, Israel). Temperatures produced by this device range from 33°C to 51°C. Participants reported subjective pain ratings to these stimuli using a computerized VAS, anchored by “No pain” and “The most imaginable pain.”

Experimental Procedures

The present study utilized a within-subjects design to assess the test-retest reliability of pain-related brain activity across three fMRI runs. Due to individual differences in pain perception, each participant completed QST during a screening visit prior to baseline fMRI scanning. Thermal pulses were delivered on the dorsal aspect of each foot, beginning at 43°C and increasing by 1°C until tolerance or 51°C was reached. Participants rated pain intensity on a VAS after each pulse. Temperature for "pain" stimuli used during the baseline fMRI visit were determined for each individual based on the lowest temperature rated between 40–60.

Scanning during the baseline fMRI visit included one anatomical and three functional MRI scans. The experimental paradigm was used for all three functional scans, and consisted of 16 thermal "pain" pulses delivered in a random order to one of four sites on the dorsal aspects of both feet. Each pulse lasted four seconds, with a 12-second interstimulus interval (Figure 1), and participants rated pain intensity following each stimulus using a computerized VAS.

Data Acquisition and Preprocessing

All MRI scanning took place on a 3.0T research-dedicated Phillips Achieva scanner, and an 8-channel head coil was used. High-resolution structural data were collected using a T1-weighted MP-RAGE protocol with the following parameters: 180 1mm sagittal slices, matrix (mm) = 256 × 256 × 180, repetition time (TR) = 8.1ms, echo time (TE) = 3.7ms, FOV (mm) = 240 × 240 × 180, FA = 8°, voxel size = $1mm^3$). Functional MRI used an echo planar acquisition protocol with the following parameters: 38 contiguous 3mm trans-axial slices, matrix (mm) = 80 × 80 × 39, TR/TE = 2000/30ms, FOV (mm) = 240 × 240 × 114, FA = 80°, voxel size = $3mm^3$). Four dummy volumes were discarded at the beginning of each fMRI run to reduce saturation effects due to B_0 field inhomogeneity. Each scan lasted five minutes and 40 seconds, and all three runs used in the present analyses were conducted consecutively.

Image preprocessing was conducted using SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK) with MATLAB 2011b (MathWorks, Sherbon, MA, USA). Functional MRI preprocessing procedures consisted of slice-time correction in ascending order, 3D motion correction with realignment to the middle volume of each sequence, and coregistration to the structural MRI. Data were then normalized to a standardized MNI template using a 4^{th} degree B-spline interpolation and spatially smoothed [6mm isotropic Gaussian kernel (FWHM)].

We examined motion parameters and signal-to-noise ratios (SNR) across all three runs to ensure that subsequent analyses were not affected by these factors. Results of a one-way ANOVA showed that runs were not significantly different for average motion [$F(2,67) = 1.266$, $p > .05$] or average SNR [$F(2,67) = 1.298$, $p > .05$], suggesting that subsequent analyses were not influenced by systematic differences in image quality.

Functional MRI Analyses and ROI Extraction

Brain activity significantly associated with thermal "pain" stimuli at the individual- and group-levels were identified using a random-effects general linear model (RFX-GLM). For individual-level contrasts, the task regressor (i.e., thermal stimulation periods) was deconvolved on the canonical HRF, and temporal/dispersion derivatives were modeled to remove confounds associated with differences in peak response latency and peak response duration, respectively. To extract values for the a priori ROIs (i.e., bilateral ACC and aINS), we conducted one-sample t-tests within inclusive anatomical masks (i.e., search spaces) created using the Automated Anatomical Labeling (AAL) atlas within the WFU PickAtlas (WFU Pickatlas, v2.4).[21] We used normalized anatomical search spaces rather than individual- or group-generated BOLD volumes of interest to emulate methods that would be potentially feasible and standardized in a clinical setting. Images from individual participants for each run were thresholded [$p < .05$, cluster minimum (k) = 5 voxels], and ROI cluster sizes and peak T-score values were extracted to calculate subsequent ICCs.

Additionally, a whole brain RFX-GLM (pain vs. no pain) was conducted on fMRI data at the group-level to examine whether expected ROI activity was in fact robust at the group-level. Data were thresholded and a priori anatomical search spaces were applied ($p_{FDR} < .05$, $k = 5$).

Test-Retest Reliability

ICCs of absolute agreement were conducted using SPSS v21.0 (SPSS Inc., Chicago, IL, USA). This statistic provides a measure of consistency through a ratio of between-subject variance to total variance,[9] and is commonly used to examine reliability of fMRI summary statistics.[3,9] Absolute agreement ICCs were conducted on ROI cluster sizes and peak T-scores, as well as VAS subjective pain intensity ratings, for the following time points: run 1 vs. run 2, run 2 vs. run 3, run 1 vs. run 3, and all runs.

Results

Individual- and Group-Level RFX-GLM

At the group-level, bilateral ACC and aINS were all significantly more activated during pain compared to no pain [L-ACC: $t(22) = 12.54$, $p_{FDR} < .001$, $k = 123$ voxels; R-ACC: $t(22) = 11.65$, $p_{FDR} < .001$, $k = 116$ voxels; L-aINS: $t(22) = 6.08$, $p_{FDR} < .05$, $k = 112$ voxels; R-aINS: $t(22) = 6.43$, $p_{FDR} < .05$, $k = 108$ voxels]. Figure 2 demonstrates brain activity associated with thermal stimuli and the group-level.

At the individual level, all participants had activity within bilateral ACC comparing pain to no pain. However, six participants lacked left aINS activity and two participants lacked right aINS activity across all three runs. Exploratory analyses via a one-samples t-test revealed that participants without identifiable aINS activity did not have significantly different pain ratings from the group average [$t(6) = -.492$, $p = .657$].

ICCs of fMRI Summary Statistics

ICC coefficients range from 0 to 1 and classification of reliability has been suggested as the following: less than 0.4 = "poor," between 0.4–0.6 = "fair," between 0.61–0.8 = "good," and greater than 0.8 = "excellent".[7,9] The results below are described in terms of these criteria.

Cluster Size Test-Retest Reliability—Table 1 shows results for ICCs on cluster size within all four ROIs. Coefficients for L-ACC ranged from 0.32 (run 1 vs. run 3) to 0.67 (run 2 vs. run 3), suggesting poor to good reliability. Among all three runs, reliability was good (ICC = 0.65). R-ACC showed somewhat better reliability, ranging from fair (run 1 vs. run 3, ICC = 0.5) to good (run 2 vs. run 3, ICC = 0.75). Average R-ACC cluster size reliability across runs was good (ICC = 0.7).

Cluster size test-retest reliability was generally higher for the insula; however, it is important to keep in mind that six and two individuals did not have activity in the L-aINS and RaINS, respectively, resulting in a cluster size of 0 for all three runs. Left aINS cluster size reliability ranged from fair (run 1 vs. run 3, ICC = 0.60) to excellent (run 2 vs. run 3, ICC = 0.83). Eight individuals lacked L-aINS activity in both runs 2 and 3. Reliability among all three runs was good (ICC = 0.79). Finally, cluster size reliability for R-aINS ranged from fair (run 1 vs. run 2, ICC = 0.47) to excellent (run 2 vs. run 3, ICC = 0.83), with overall good reliability (ICC = 0.73). Three individuals lacked R-aINS activity in both runs 2 and 3.

Peak T-score Test-Retest Reliability—Peak t-scores for the L-ACC showed good reliability between all run pairs and across all three runs (Table 2). Right ACC peak t-score reliability ranged from good (run 1 vs. run 3, ICC = 0.63) to excellent (run 2 vs. run 3, ICC = 0.83), with overall good reliability among the three runs (ICC = 0.8). Again, test-retest reliability was highest for aINS activity, which ranged from good to excellent for both ROIs (Table 2). Among all three runs, aINS peak t-score reliability was excellent for left (ICC = 0.87) and right (ICC = 0.86) regions.

ICCs of VAS Subjective Pain Ratings

All participants reported pain related to thermal stimulation during fMRI scanning (mean VAS score = 40.27; SD = 15.7), including participants who lacked aINS activity. Test-retest reliability of VAS pain ratings was excellent between all run comparisons, as well as among all three runs (ICC range = .926 – .958). Table 3 shows the ICC coefficients for pain ratings among all runs.

Discussion

The present study examined the test-retest reliability of the anterior cingulate cortex (ACC) and anterior insula (aINS) in healthy controls undergoing experimental pain. Across three consecutive fMRI runs, intraclass correlation (ICC) coefficients were generally classified as having "good" reliability for both ROI cluster size and peak T-score. Bilateral aINS reliability was superior to bilateral ACC reliability; however, because six and two individuals completed lacked left and right aINS activity across all three runs, respectively, this ICC coefficient is potentially artificially inflated. The absent aINS activity in these

individuals is a departure from findings at the group-level, which showed robust activity associated with painful thermal stimuli within all four ROIs. Overall, subjective report of pain via Visual Analogue Scales (VASs) showed stronger test-retest reliability compared to our method of extracting and comparing fMRI summary statistics, with all ICC coefficients falling within the “excellent” range.

Consistency of Present Results with Previous Studies

Compared to previous studies of fMRI test-retest reliability, our fMRI results aligned with or exceeded findings across other fields. As previously described, Bennett and Miller[3] generally found fMRI test-retest reliability ranging from “fair” to “good” across studies included in their meta-analysis, which reported on various methods of processing and extracting data for reliability analyses. Although most fMRI studies of reliability examine group-level findings, it is important to measure this psychometric property at the individual-level for potential clinical translation. Plichta and colleagues[25] showed “excellent” test-retest reliability for group-level data across three cognitive tasks; however, ICC coefficients for within-subject reliability ranged from “fair” to “good.” Our results of individual-level reliability aligned with this finding, as ICCs generally fell within the “good” range.

Similarly, our test-retest reliability results for subjective pain report were on the magnitude of previous studies. Specifically related to the VAS, previous studies have shown ICC coefficients within the “excellent” range.[4,15,34] In addition to their high test-retest reliability, pain VASs have been found to have other excellent psychometric properties including a ratio scale level of measurement,[26] capacity to discriminate very small differences in intensity,[27] and capacity to measure multiple pain dimensions.[26,27] The extent to which such properties apply to brain imaging variables remains to be determined.

Taken together, these results demonstrate that self-report of pain has generally been shown to have relatively higher test-retest reliability compared to methods used in the reported fMRI findings. They also suggest that poor reliability of self-report is unsupportable, and certainly not a reasonable rationale for seeking a brain-based biomarker of pain.

Standardization Issues

Several important issues of standardization within the clinical neuroimaging field are important to acknowledge for the interpretation of the present data. First, methods for fMRI data collection, preprocessing, and statistics vary widely. Therefore, it is possible that our test-retest reliability results, as well as those described in the meta-analysis of fMRI reliability,[3] were influenced by the analysis decisions made, rather than the technology itself. Standardization of echo planar imaging protocols and analysis pipelines might be helpful for potential translation of this technology to the clinic.

Another important question across the field of clinical neuroimaging is the acceptability criterion of test-retest reliability for the use of fMRI in clinical practice.[3,22,29] Several qualitative descriptors have been suggested to categorize ICC coefficients,[2,9,10,14] but these descriptors diverge in what constitutes “poor,” “fair,” “good,” or “excellent” reliability. This issue is important to establish for standardization across future studies of test-retest reliability in pain neuroimaging, so that results can be consistently compared.

Additionally, reliability metrics are widely unstandardized across fMRI studies, and no conclusive "optimal" approach to fMRI reliability testing has been determined. For example, some studies use ICC of fMRI summary statistics, whereas other studies calculate reproducibility ratios of voxel overlap. Because studies examining individual-level data have shown high variability in peak ROI location, this metric is less frequently used.[29] Our method of using anatomical search spaces on data analyzed via a general linear model showed generally "good" reliability; however, other techniques might yield improved reliability. Another potential metric could be the test-retest reliability of functional connectivity between pain-related ROIs, rather than fMRI summary statistics. Functional connectivity between pain-related regions might have higher specificity as a measure of pain.

Potential for Clinical Use

Although the present fMRI summary statistics showed generally good test-retest reliability, they did not outperform reliability of subjective pain report in this study. For fMRI to act as a surrogate of self-report due to "unreliability" of subjective measures,[20,33] fMRI reliability should exceed values demonstrated by patient report. Averaging runs tended to improve reliability, suggesting that this approach might be optimal to using single-run data. Of note, some individuals lacked ROI activity with our approach that survived standardized thresholding; however, their report of pain was not significantly different from individuals who did show activity in these regions. This finding emphasizes the possibility of individuals reporting pain, but with an "objective" biomarker that suggests otherwise.

Although we believe the present reliability results suggest that fMRI should not be used as a surrogate for self-report, we continue to emphasize the value of this technology for understanding mechanisms of acute and chronic pain, as well as helping to advance treatments for these conditions. In fact, our current reliability results suggest that one potential avenue for use of this tool in a clinical setting could be in aiding individualized treatment planning. For example, a recent study showed that certain patterns of PET activity in the insula predicted responsiveness to either cognitive-behavioral therapy or medication for treatment of depression.[24] Focusing efforts on fMRI for more efficient treatment optimization, rather than diagnosis, could potentially minimize the frustration, time, and money of using a non-individualized approach to pain management, without possibly denying treatment to a patient who reports pain without a concomitant neural biomarker. However, more research is needed to determine to what extent fMRI can be used to help optimize individual treatments.

Strengths and Limitations

To the best of our knowledge, this study is one of the first studies to report the test-retest reliability of fMRI summary statistics in acute pain processing. Rather than examining the reliability of group-level findings over time, our approach focused on individual-level reproducibility to assess the feasibility of fMRI as a potential diagnostic tool. Additionally, we compared this information with participants' subjective pain report to gauge whether fMRI outperformed test-retest reliability of self-report, as suggested in several recent reviews.[1,5,6,32,33]

However, it is important to acknowledge the present study's limitations. First, we used a sample of healthy controls, so our results reflect test-retest reliability of acute pain. Future studies should examine test-retest reliability of fMRI for chronic pain conditions. Second, we measured reliability using one analytic approach to generate ROIs. Future studies should also examine whether other methodological approaches yield better test-retest reliability. These potential approaches could examine different types of behavioral tasks (e.g., resting-state versus goal-directed task), individual-level analyses (e.g., functional connectivity versus effective connectivity, alternative ROIs), and additional functional or structural imaging modalities. Third, we examined reliability over scans within the same day, so cannot generalize about test-retest reliability over longer spans of time. Finally, data included were collected on one scanner. Because several multi-center studies of fMRI test-retest reliability have shown generally poor results, it is important to assess the reliability of findings across different sites for this tool to be used in clinical practice.

Conclusion

Many factors can influence a person's experience and subjective report of pain, such as assessment setting, relationship with the provider, mood, and motivational factors.[19] It is important to remember that neuronal activity is not immune to these same variables, and an exact set of brain regions has not yet been shown to consistently represent the presence of pain across studies.[1] Our present results suggest generally good reliability of fMRI summary statistics using our methods, highlighting the value of this technology in providing meaningful information about the etiology of chronic pain conditions and a better understanding of treatment mechanisms. However, the current test-retest reliability findings suggest that this technology does not exceed self-report of pain under the current methods described, which presents a potential limitation for the translation of fMRI as a single-subject diagnostic tool. Additionally, this finding suggests that fMRI is not yet up to the standards needed for clinical implementation that we would expect from other biomarkers. As emphasized by several authors, pain is a subjective experience, and its diagnosis should continue to rely on the patient's report.[11,12,23,28,31,34]

Acknowledgments

Supported by: Grants from the National Center for Complementary and Alternative Medicine to Dr. Michael E. Robinson (R01AT001424-05A2) and Janelle E. Letzen (F31AT007898-01A1)

References

1. Apkarian AV, Hashmi JA, Baliki MN. Pain and the brain: Specificity and plasticity of the brain in clinical chronic pain. Pain. 2011; 152
2. Aron AR, Gluck MA, Poldrack RA. Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage. 2006; 29:1000–1006. [PubMed: 16139527]
3. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. 2010; 1191:133–155. [PubMed: 20392279]
4. Bijur PE, Silver W, Gallagher EJ. Reliability of the visual analog scale for measurement of acute pain. Acad Emerg Med. 2001; 8:1153–1157. [PubMed: 11733293]
5. Borsook D, Becerra L, Hargreaves R. Biomarkers for chronic pain and analgesia. Part 1: the need, reality, challenges, and solutions. Discov Med. 2011a; 11:197–207. [PubMed: 21447279]

6. Borsook D, Becerra L, Hargreaves R. Biomarkers for chronic pain and analgesia. Part 2: how, where, and what to look for using functional imaging. Discov Med. 2011b; 11:209–219. [PubMed: 21447280]
7. Brandt DJ, Sommer J, Krach S, Bedenbender J, Kircher T, Paulus FM, Jansen A. Test-retest reliability of fMRI brain activity during memory encoding. Front Psychiatry. 2013; 4:163. [PubMed: 24367338]
8. Brooks JC, Nurmikko TJ, Bimson WE, Singh KD, Roberts N. fMRI of thermal pain: Effects of stimulus laterality and attention. Neuroimage. 2002; 15:293–301. [PubMed: 11798266]
9. Caceres A, Hall DL, Zelaya FO, Williams SCR, Mehta MA. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage. 2009; 45:758–768. [PubMed: 19166942]
10. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. Am J Ment Defic. 1981; 86:127–37. [PubMed: 7315877]
11. Coghill RC, McHaffie JG, Yen YF. Neural correlates of interindividual differences in the subjective experience of pain. Proc Natl Acad Sci U S A. 2003; 100:8538–8542. [PubMed: 12824463]
12. Davis KD. Neuroimaging of pain: what does it tell us? Curr Opin Support Palliat Care. 2011; 5:116–121. [PubMed: 21415755]
13. Duerden EG, Albanses MC. Localization of pain-related brain activation: A meta-analysis of neuroimaging data. Hum Brain Mapp. 2013; 34:109–149. [PubMed: 22131304]
14. Eaton KP, Szaflarski JP, Altaye M, Ball AL, Kissela BM, Banks C, Holland SK. Reliability of fMRI for studies of language in post-stroke aphasia subjects. Neuroimage. 2008; 41:311–322. [PubMed: 18411061]
15. Gallagher EJ, Bijur PE, Latimer C, Silver W. Reliability and validity of a visual analog scale for acute abdominal pain in the ED. Am J Emerg Med. 2002; 20:287–290. [PubMed: 12098173]
16. Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: Assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. Clin J Pain. 2005; 21:73–82. [PubMed: 15599134]
17. Jacob T, Baras M, Zeev A, Epstein L. Low back pain: Reliability of a set of pain measurement tools. Arch Phys Med Rehabil. 2001; 82:735–742. [PubMed: 11387576]
18. Jones DH, Kilgour RD, Comtois AS. Test-retest reliability of pressure pain threshold measurements of the upper limb and torso in young healthy women. J Pain. 2007; 8:650–656. [PubMed: 17553750]
19. Jensen MP. The validity and reliability of pain measures in adults with cancer. J Pain. 2003; 4:2–21. [PubMed: 14622723]
20. Mackey SC. Central neuroimaging of pain. J Pain. 2013; 14:328. [PubMed: 23548485]
21. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fmri data sets. NeuroImage. 2003; 19:1233–1239. [PubMed: 12880848]
22. Meindl T, Teipel S, Elmouden R, Mueller S, Koch W, Dietrich O, Coates U, Reiser M, Glaser C. Test–retest reproducibility of the default-mode network in healthy individuals. Hum Brain Mapp. 2011; 31:237–246. [PubMed: 19621371]
23. McCaffrey M, Beebe A. Giving narcotics for pain. Nursing. 1989; 19:161–165. [PubMed: 2573871]
24. McGrath CL, Kelley ME, Holtzheimer PE, Dunlop BW, Craighead WE, Franco AR, Craddock RC, Mayberg HS. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA Psychiatry. 2013; 12:1–9.
25. Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes ABM, Sauer C, Tost H, Esslinger C, Colman P, Wilson F, Kirsch P, Meyer-Lindenberg A. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage. 2012; 60:1746–1758. [PubMed: 22330316]
26. Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain. 1983; 17:45–60. [PubMed: 6226917]

27. Price DD, Patel R, Robinson ME, Staud R. Characteristics of electronic visual analogue and numerical scales for ratings of experimental pain in healthy subjects and fibromyalga patients. Pain. 2008; 140:158–166. [PubMed: 18786761]

28. Robinson ME, Staud R, Price DD. Pain measurement and brain activity: Will neuroimages replace pain ratings? J Pain. 2013; 14:323–327. [PubMed: 23548484]

29. Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Scheltens P. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn Reson Imaging. 1998; 16:105–113. [PubMed: 9508267]

30. Staud R. Predictors of clinical pain intensity in patients with fibromyalgia syndrome. Curr Pain Headache Rep. 2005; 9:316–21. [PubMed: 16157059]

31. Sullivan MD, Cahana A, Derbyshire S, Loeser JD. What does it mean to call chronic pain a brain disease? J Pain. 2013; 14:317–322. [PubMed: 23548483]

32. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-based neurologic signature of physical pain. N Engl J Med. 2013; 368:1388–1397. [PubMed: 23574118]

33. Wartolowska K. How neuroimaging can help us to visualise and quantify pain? Eur J Pain Suppl. 2011; 5:323–327.

34. Williamson A, Hoggart B. Pain: A review of three commonly used pain rating scales. J Clin Nurs. 2005; 14:798–804. [PubMed: 16000093]

Perspective

This study is one of the first reports to demonstrate the test-retest reliability of fMRI findings related to pain processing, and provides a comparison to the reliability of subjective reports of pain. This information is essential for determining whether fMRI technology should be potentially translated for clinical use.

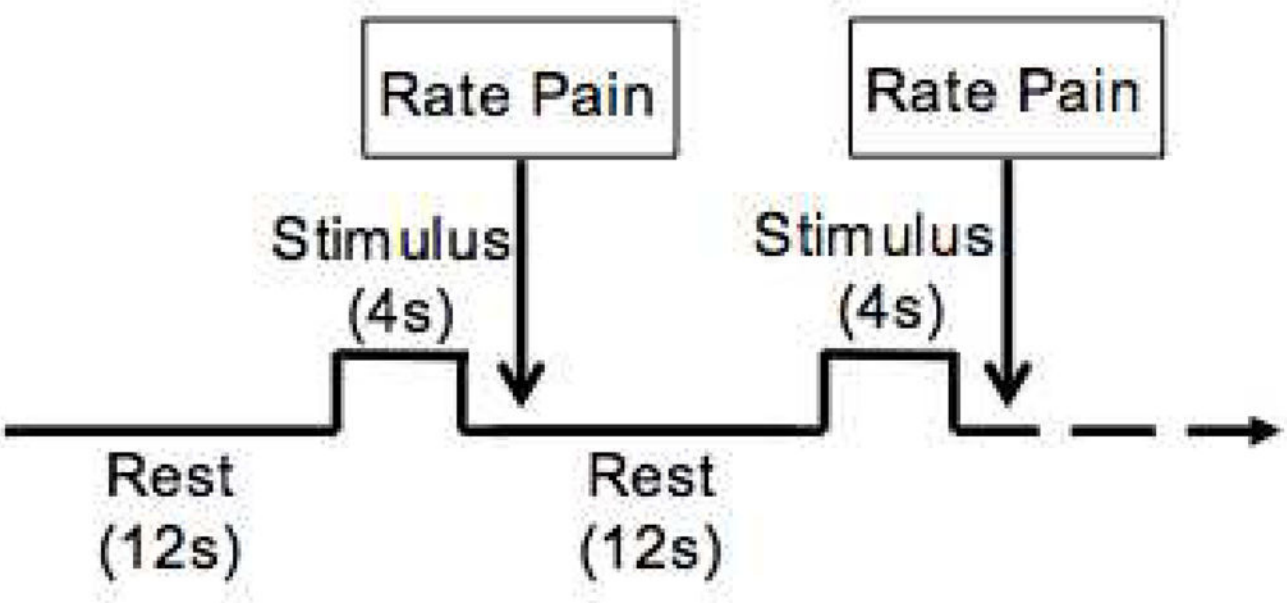

Figure 1.

Three fMRI runs were collected. The paradigm for each run included 16 4-second thermal pulses with a 12-second interstimulus interval. The temperature used for the thermal pulses was specific to each participant, based on quantitative sensory testing prior to scanning. Participants rated pain intensity using a computerized VAS subsequent to each stimulus.

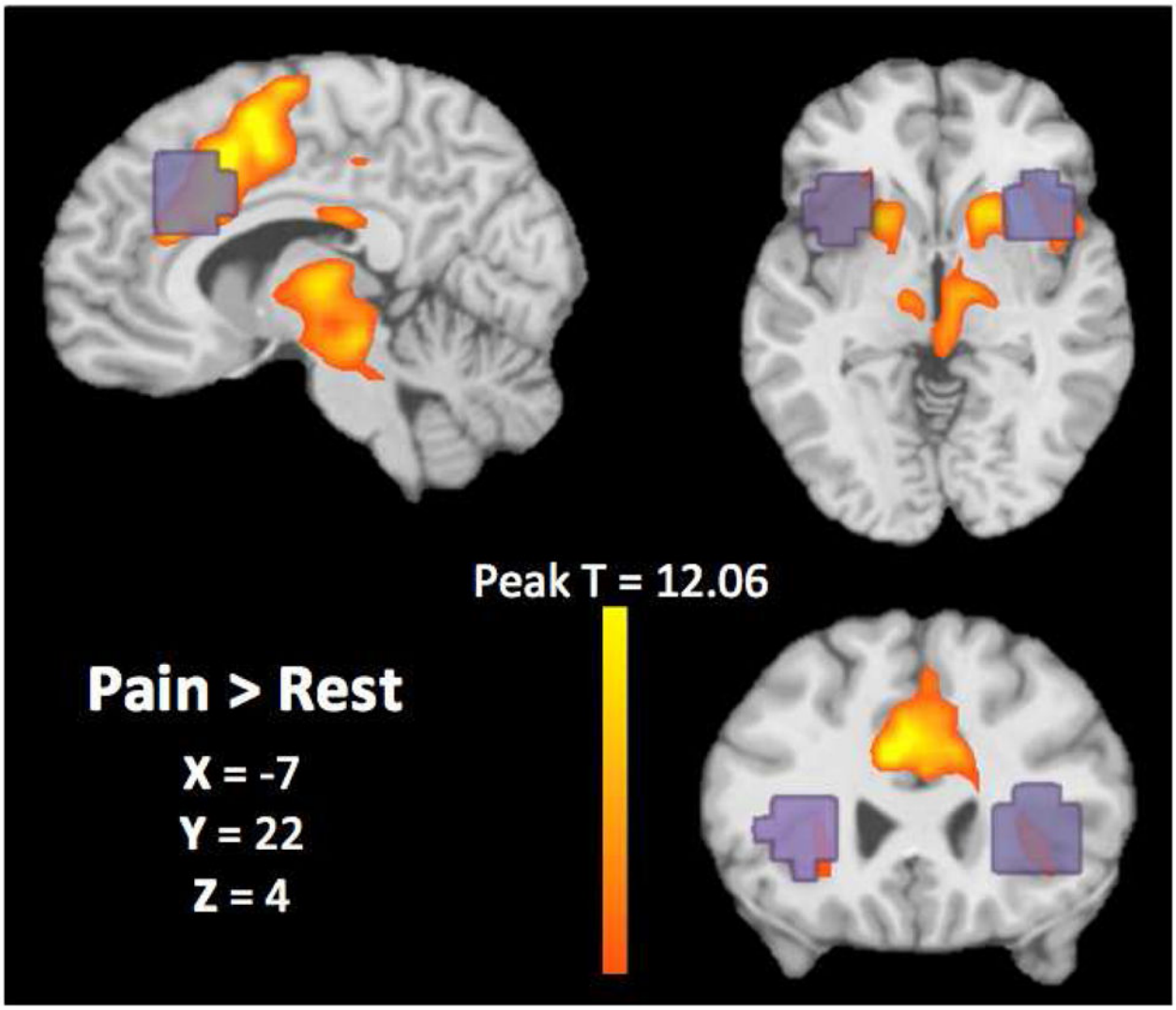

Figure 2.

A group-level contrast was calculated to examine brain activity associated with painful thermal stimuli compared to no stimulus (orange). Results showed significant activity within our a priori ROIs of bilateral ACC and aINS ($p_{FDR} < .001$ and $p_{FDR} < .05$, respectively). Anatomical ROIs used to extract individual-level fMRI summary statistics are overlaid on the activation map in purple.

Table 1

Intraclass correlation coefficients for cluster size test-retest reliability among fMRI runs

ROI	Run 1 vs. Run 2	Run 2 vs. Run 3	Run 1 vs. Run 3	All Runs
Left ACC	.654**	.665**	.315	.653**
Right ACC	.554*	.750**	.499*	.703***
Left aINS	.718**	.831***	.595*	.791***
Right aINS	.473	.830***	.573*	.732***

* significant at $p < .05$;

** significant at $p < .01$;

*** significant at $p < .001$

Region of Interest (ROI), Anterior Cingulate Cortex (ACC), anterior Insula (aINS)

Table 2

Intraclass correlation coefficients for ROI peak T-score test-retest reliability among fMRI runs

ROI	Run 1 vs. Run 2	Run 2 vs. Run 3	Run 1 vs. Run 3	All Runs
Left ACC	.611*	.707***	.637**	.749***
Right ACC	.663**	.827***	.636**	.795***
Left aINS	.849***	.763***	.787***	.859***
Right aINS	.768**	.884***	.792***	.870***

* significant at $p < .05$;

** significant at $p < .01$;

*** significant at $p < .001$

Region of Interest (ROI), Anterior Cingulate Cortex (ACC), anterior Insula (aINS)

Table 3

Intraclass correlation coefficients for VAS pain ratings test-retest reliability among fMRI runs

Run 1 vs. Run 2	Run 2 vs. Run 3	Run 1 vs. Run 3	All Runs
.926***	.943***	.946***	.958***

* significant at $p < .05$;

** significant at $p < .01$;

*** significant at $p < .001$