**Manuscript outline:**

**Title: Effect size and reliability of the Neurological Pain Signature**

**Methods:**

**Participants:**

The data used for this study are based on a single trial database on healthy subjects during pain tasks including comprehensive behavioral and fMRI data. Our data set included overall ∼17,000 single-trial images of fMRI activity associated with multiple levels of noxious heat and pain ratings, across over health 320 participants from 9 studies. Descriptive data on age, sex and other features of each study sample are given in **Table 1**. In all studies, participants received a series of contact-heat stimuli and rated their individually experienced pain following each stimulus. The number of trials, stimulation sites, rating scales and stimulus intensities and durations varied across studies but were comparable; these variables are summarized in **Table 2**. Each study also comprises a specific psychological manipulation, such as cue-induced expectation and placebo treatment. In the studies included, we focused only on residual pain ratings (ratings after removing noxious stimulus intensity) and the NPS irrespective of the study-specific psychological manipulations.

All data from the studies have been used in previous publications (see **Table 1** and ref. Lindquist et al. 2015 and Woo et al. 2017). However, the analyses and findings reported here are novel and have not been published elsewhere. All participants were recruited from New York City and Boulder/Denver Metro Areas. The institutional review board of Columbia University and the University of Colorado Boulder approved all the studies, and all participants provided written informed consent. Preliminary eligibility of participants was determined through an online questionnaire, a pain safety screening form and an MRI safety screening form. Participants with psychiatric, physiological or pain disorders, neurological conditions, and MRI contraindications were excluded prior to enrollment. No participants were excluded from the study after screening other than individuals who, upon screening, provided different responses that made them now ineligible (eg, development of a physiological disorder).

>>> *Table 1* “ Study demographics and prior publications ” <<

**Materials and Procedures**

**Thermal stimulation**

In each study, we delivered thermal stimulation to multiple skin sites using a TSA-II Neurosensory Analyzer (Medoc Ltd., Chapel Hill, NC) with a 16 mm Peltier thermode endplate (Study SCEBL: 32 mm). On every trial, after the offset of stimulation, participants rated the magnitude of the warmth or pain they had felt during the trial on a visual analog scale or labelled magnitude scale. Other thermal stimulation parameters varied across studies, with stimulation temperatures ranging from 41 °C to 50 °C\*\*\* and stimulation durations from 1,85 to 16 seconds\*\*\*. Most studies applied thermal stimulation to the glabrous skin of the left forearm; Study BMRK 4 (additionally to the dorsum of the left foot), IE2(lower leg only) and SCEBL (leg only) applied the stimulation to the lower extremity. See **Table 2** for stimulation duration, intensity levels and location of stimulus site, details of the rating scales, duration of inter-stimulus interval, and number of trials per subject. .

>>> *Table 2* “Pain stimuli information” <<

**fMRI Analysis**

**Preprocessing**.

Structural T1-weighted images were co-registered to the mean functional image for each subject using the iterative mutual information-based algorithm implemented in SPM (Ashburner and Friston, 2005), and were then normalized to MNI space using SPM. Following SPM normalization, Studies 4 (nsf) and study 6 (exp) included an additional step of normalization to the group mean using a genetic algorithm-based normalization (Atlas et al., 2010, 2014; Wager and Nichols, 2003). In each functional dataset, we removed initial volumes to allow for image intensity stabilization (see **Table 3** for details on acquisition and preprocessing parameters). We also identified image-intensity outliers (that is, ‘spikes’) by computing the mean and standard deviations (SD, across voxels) of intensity values for each image for all slices to remove intermittent gradient and severe motion-related artefacts present to some degree in all fMRI data. To identify outliers, we first computed both the mean and the S.D. of intensity values across each slice, for each image. Mahalanobis distances for the matrix of (concatenated) slice-wise mean and standard deviation values by functional volumes (over time) were computed. Any values with a significant χ2 value (corrected for multiple comparisons based on the more stringent of either false discovery rate or Bonferroni methods) were considered outliers. In practice, less than 1% of images were deemed outliers. The outputs of this procedure were later included as nuisance covariates in the first level models. Next, functional images were corrected for differences in the acquisition timing of each slice and were motion-corrected (realigned) using SPM. The functional images were warped to SPM's normative atlas (warping parameters estimated from co-registered, high-resolution structural images), interpolated to 2 × 2 × 2 mm3 voxels, and smoothed with an 8 mm FWHM Gaussian kernel. This smoothing level has been shown to improve inter-subject functional alignment, while retaining sensitivity to mesoscopic activity patterns that are consistent across individuals (Shmuel et al. 2010).

>>> *Table 3* “Imaging Acquisition parameters” <<<

**Single trial analyses**

Single trial analysis (Except Study 2/bmrk4 and Study 4/exp). For each study a single trial, or “single-epoch”, design and analysis approach was employed to model the data. Quantification of single trial response magnitudes was done by constructing a GLM design matrix with separate regressors for each trial, as in the “beta series” approach (Mumford et al., 2012; Rissman et al., 2004). First, boxcar regressors, convolved with the canonical hemodynamic response function (HRF), were constructed to model cue, pain, and rating periods in each study. Then, we included a regressor for each trial, as well as several types of nuisance covariates. Because each trial consisted of relatively few volumes, trial estimates could be strongly affected by acquisition artifacts that occur during that trial (e.g. sudden motion, scanner pulse artifacts, etc.). Therefore, trial-by-trial variance inflation factors (VIFs; a measure of design-induced uncertainty due, in this case, to collinearity with nuisance regressors) were calculated, and any trials with VIFs that exceeded 2.5 were excluded from the analyses. For Study 8 (nsf), we also excluded global outliers (trials that exceeded three SDs above the mean), and employed a principal component based denoising-step during preprocessing to minimize artifacts. This approach generated single trial estimates that reflect the amplitude of the fitted HRF on each trial and refer to the magnitude of anticipatory and pain-period activity for each trial in each voxel. Single trial analysis for Study 2 (bmrk4) and Study 4 (exp) were based on fitting a set of three basis functions, rather than the standard HRF used in the other studies. This flexible strategy allowed the shape of the modeled hemodynamic response function (HRF) to vary across trials and voxels. This procedure differed from that used in other studies because (a) it maintains consistency with the procedures used in the original publication on Study 6 (exp) (Atlas et al., 2010), and (b) it provides an opportunity to examine predictive performance using a flexible basis set. For both studies, the pain period basis set consisted of three curves shifted in time and was customized for thermal pain responses based on previous studies (Atlas et al., 2010; Lindquist et al., 2009). To estimate cue-evoked responses for Study 6, the pain anticipation period was modeled using a boxcar epoch convolved with a canonical HRF. This epoch was truncated at 8 s to ensure that fitted anticipatory responses were not affected by noxious stimulus-evoked activity. As with the other studies, we included nuisance covariates and excluded trials with VIFs > 2.5. In Study 6 we also excluded trials that were global outliers (those that exceeded 3 SDs above the mean). We reconstructed the fitted basis functions from the flexible single trial approach to compute the area under the curve (AUC) for each trial and in each voxel. We used these trial-by-trial AUC values as estimates of trial-level anticipatory or pain-period activity.

**Computing Neurologic Pain Signature (NPS) responses.**

We computed for each trial and each subject a single scalar value representing their expression of the NPS pattern in response to the thermal pain stimulus (using the contrast [Pain Stimulation minus Baseline] images as detailed below). There are mainly three methods to calculate the NPS pattern response: (1) dot-product (NPSdot = ) which combine magnitude and spatial similarity information; (2) correlation (NPScorr = ) which excludes information related to whole-image mean and scale, and represents spatial pattern similarity; (3) cosine similarity (NPScos = ), which is similar with correlation, but without mean-centering. We reported the major results using NPSdot in the main text and compared the effect size and reliability of three versions as well (see **Table S4**).

To test performance of brain regions within the NPS and whether NPS's performance exceeds that of individual brain regions, we also computed the pattern expression for each brain area within NPS. We compared the effect size and the reliability of the individual brain regions and the NPS. In most of the regions in the NPS, pain is associated with increased overall activity. Such regions include the major targets of ascending nociceptive afferents, including the thalamus, secondary somatosensory regions (SI/SII), posterior, mid and anterior insula and adjacent opercula, midbrain, dorsal anterior cingulate cortex (dACC), inferior frontal gyrus and amygdala (REF). In a subset of other medial regions, including the perigenual ACC (pgACC) and the PCC (posterior cingulate)/precuneus/paracentral lobule, pain was associated with deactivation in the original NPS pattern. These regions are not strongly linked to nociception and are not direct targets of nociceptive afferents, rather they have been associated with a variety of affective, autonomic, social, self-referential, and decision-making functions [Roy et al\*\*\*].

**Results:**

**Four types of NPS effect size**

We evaluated the NPS response to thermal pain stimulus from four tests. Firstly, NPS responses in the contrast of [Pain minus Baseline] were significantly larger than zero in each of 9 studies (t = 5.02 – 19.05, ps < 0.001, d = 1.22 – 2.62; see **Figure S1(A)** for the mean response of NPS of each participant and **Figure 1(A)** for the effect size; see **Table S1** for the statistic details of each study). Secondly, the correlations of NPS with temperature within each subject were significantly larger than zero in each of 9 studies ( within-person-r = 0.05 – 0.42, t = 2.32 – 18.91, ps < 0.05, d = 0.53 – 2.74; see **Figure S1(B)** for the within-subject correlation between NPS and temperature and **Figure 1(A)** for the effect size; see **Table S1** for the statistic details of each study). Thirdly, the correlation of NPS with subjective pain rating within each participant were significantly larger than zero in each of 9 studies (within-person-r = 0.14 – 0.42, t = 4.81 – 11.49, ps < 0.001, d = 0.94 – 2.13; see **Figure S1(C)** for the within-subject correlation between NPS and temperature and **Figure 1(A)** for the effect size; see **Table S1** for the statistic details of each study). Lastly, the correlations of NPS with subjective pain rating between-participant (i.e., individual differences) were only significant in 2 out of 9 studies (between-person-r = 0.13 – 0.74, p = 0.69e-3 – 0.70, d = -0.27 – 2.20; see Figure 1(D) for the between-subject correlation between NPS and subjective pain rating and Figure 2(A) for the effect size; see Table S1 for the statistic details of each study).

We did the same analyses for each local brain area of NPS and compared the effect sizes of single brain areas and the NPS. Generally, positive brain regions had higher effect size than the negative brain regions, and the effect sizes of NPS were the highest in all four tests (see **Figure 1(A)**, **Figure S2** and **Table S2** for details). We also conducted a paired t-test between the effect size of NPS and the effect size of each local area. For the effect size of mean response, NPS was not significantly different from rIns (mean±se = 1.94±0.14 vs. 1.69±0.17); for the effect size of within-subject correlation with the temperature, NPS was not significantly different from rIns (1.64±0.27 vs. 1.24±0.23), dACC (1.14±0.19), lIns (0.97±0.16), and rS2 (0.78±0.13); for the effect size of within-subject correlation with the subjective pain ratings, NPS was not significantly different from dACC (1.52±0.15 vs. 1.20±0.11), rIns (1.15±0.13); and for the effect size of between-subject correlation with the subjective pain ratings, NPS was not significantly different from any brain regions.

**Reliability of NPS**

NPS response was extracted for each trial of each participant. The mean NPS responses were calculated for the first half and second half trials for each participant. Reliability was quantified using a two-way mixed effects ICC with time (1st vs. 2nd half) as a fixed effect, and subject as a random effect. The mixed-effect model is referred to as ICC(3,k) according to McGraw and Wong (1996) and defined as ICC (3,k) = (BMS - EMS) / BMS (Shrout and Fleiss, 1979), where BMS represents between-subjects mean square, and EMS represents error mean square. ICC tracks the consistency of measures and is commonly used in studies of task-fMRI test-retest reliability. To test reliability for each local region of NPS, we calculated ICCs for the pattern response of each local region as well. To compare reliability of brain pattern and subjective pain rating, we also calculated ICC of subjective pain rating.

Reliability of NPS was distributed from good to excellent among 9 datasets (mean±s.e. of ICC = 0.83±0.03), which was significantly smaller than the reliability of subjective pain rating (0.91±0.02, t(8) = 4.71, p = 0.002). Reliability of each local brain region was not larger than the NPS (ICC = 0.54 – 0.80). Among them, reliability of rThal and pgACC were significantly smaller than NPS (see **Figure 1(B)** and **Table S3** for statistical details).

*How observation numbers influence reliability?*

In task-fMRI studies, we usually measured the brain response by averaging multiple trials of a condition. Here, we tested how trial numbers of heat stimuli might influence the reliability of NPS. The results in **Figure 2(A)** left panel showed that the more trials used to calculate the NPS response, the higher reliability we got in all 9 datasets. This was also true for the reliability of subjective pain rating (the right panel). On average, 30 or more trials were required to get an excellent reliability of NPS. The reliability of subjective pain rating was overall very high with even only one trial.

*How does the time interval between sessions influence reliability?*

In the 9 single-trial datasets, we divided all trials collected within one session into first half and second half and calculated the ICC of two measures. To further test how the time interval between sessions influenced the reliability, we included another two datasets, RTNF and OLP4CBP, which had multiple sessions collected for each participant with a time interval from days to months. RTNF included 27 healthy participants who received the painful stimuli in three sessions with 1-week interval on average in between. In each session, there are 30 trials of heat stimuli applied on the lower leg. OLP4CBP includes 40 participants with chronic back pain. Participants in this study received the painful stimuli in two sessions with 1-month interval. In each session, there are 5 trials of pressure pain applied on the thumb. The results in **Figure 2(A)** left panel showed that the NPS in these two datasets with longer time intervals showed comparable reliability with single-trial datasets under the same observation number. The results in **Figure 2(A)** right panel showed the RTNF showed comparable reliability, while OLP4CBP showed poorer reliability in subjective pain rating compared with single-trial dataset. Probably because the participants in OLP4CBP were with chronic pain and showed hypersensitivity in pain reports (ref \*\*\*), which might lead to low reliability in subjective pain rating.

*How does the size of manipulation influence reliability?*

The property of the stimulus itself might influence the reliability, such as the effect size it induced. Heat stimuli with higher temperature could generally induce higher pain effects. The results in **Figure 2(B)** left panel showed that NPS responses induced by higher temperature had higher reliability. However, this was not true for the subjective pain rating, which was very reliable across all temperatures. NPS response might be more specific for the high pain stimuli, while subjective pain rating could represent different pain levels in a reliable way.

*How contrast calculation influences reliability?*

There are two commonly used methods to calculate the brain response of a condition, e.g., comparing with the baseline or a control condition. The reliability of the control condition might greatly influence the reliability of the interest. The results in **Figure 2 (C)** left panel showed that reliability of NPS dropped when the response of NPS was calculated by the contrast between a high and a low temperature (i.e., the control condition). The drop of the reliability was smaller in subjective pain rating since the subjective pain rating was reliable across all temperatures.