Research Paper

PAIN 00 (2024) 1–16

PAIN®

OPEN

Decoding pain: uncovering the factors that affect the performance of neuroimaging-based pain models

Dong Hee Lee[a,b,c], Sungwoo Lee[a,b,c], Choong-Wan Woo[a,b,c,*]

Abstract

Neuroimaging-based pain biomarkers, when combined with machine learning techniques, have demonstrated potential in decoding pain intensity and diagnosing clinical pain conditions. However, a systematic evaluation of how different modeling options affect model performance remains unexplored. This study presents the results from a comprehensive literature survey and benchmark analysis. We conducted a survey of 57 previously published articles that included neuroimaging-based predictive modeling of pain, comparing classification and prediction performance based on the following modeling variables—the levels of data, spatial scales, idiographic vs population models, and sample sizes. The findings revealed a preference for population-level modeling with brain-wide features, aligning with the goal of clinical translation of neuroimaging biomarkers. However, a systematic evaluation of the influence of different modeling options was hindered by a limited number of independent test results. This prompted us to conduct benchmark analyses using a locally collected functional magnetic resonance imaging dataset (N = 124) involving an experimental thermal pain task. The results demonstrated that data levels, spatial scales, and sample sizes significantly impact model performance. Specifically, incorporating more pain-related brain regions, increasing sample sizes, and averaging less data during training and more data during testing improved performance. These findings offer useful guidance for developing neuroimaging-based biomarkers, underscoring the importance of strategic selection of modeling approaches to build better-performing neuroimaging pain biomarkers. However, the generalizability of these findings to clinical pain requires further investigation.

Keywords: Neuroimaging, Predictive modeling, Biomarker, Classification, Machine learning

1. Introduction

Neuroimaging-based biomarkers of pain are increasingly gaining attention in basic and clinical studies of pain.[9,35] In clinical settings, pain assessment primarily relies on self-report, which is considered the gold standard.[42] However, self-report may be limited in capturing the complex biopsychosocial nature of pain in some contexts.[3,21,39] To address these limitations and provide a more comprehensive assessment, complementary measures have been explored.[34] Neuroimaging data provide a unique window that allows us to assess pain based on brain structure and functions, having the potential to serve as biomarkers for the prediction of pain intensity and the diagnosis of clinical pain conditions. According to the FDA-NIH Biomarker Working Group's Biomarkers, EndpointS, and other Tools Resource, biomarkers can be categorized into multiple types, each with its own clinical utility (Table S1, http://links.lww.com/PAIN/C129). These include diagnostic, predictive, prognostic, safety, pharmacodynamic/response, monitoring, and susceptibility/risk biomarkers.[11] In this study, we investigated how modeling targets and options influence the performances of neuroimaging pain biomarkers through a systematic literature survey and benchmark analyses (**Fig. 1**).

We compared model performances focusing on different modeling targets, data levels, spatial scales, model levels, and sample sizes. These variables were chosen based on their significance in prior research. For example, owing to a high noise level in neuroimaging data, researchers usually average data across multiple trials to enhance the signal-to-noise ratio. However, it is unclear whether data averaging is always beneficial to model performance. While machine learning algorithms generally necessitate substantial data to recognize meaningful patterns hidden in the data, data averaging decreases the quantity of data, potentially along with informative variances, in favor of an improved signal-to-noise ratio. Previous studies commonly suggested that increased data averaging improves classification accuracy and explained variance.[22,43] However,

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea, [b] Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea, [c] Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, South Korea

**Corresponding author. Address: Department of Biomedical Engineering, Sungkyunkwan University, Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon 16419, Republic of Korea. Tel.: +82 (31) 299-4363. E-mail address: waniwoo@skku.edu (C.-W. Woo).*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.painjournalonline.com).

http://dx.doi.org/10.1097/j.pain.0000000000003392

these studies typically averaged both training and testing data to ensure consistent data distribution across training and testing sets. Consequently, the impact of data averaging on the model generalizability, particularly when the test data have a different data distribution, requires further investigation.

Furthermore, it is well known that the brain representations of pain are distributed across multiple brain systems.[7,8] Consistent with this idea, previous studies have shown that predictive models containing more voxels and regions show better model performance.[4,18,25] However, it is unclear whether including more voxels and regions is always beneficial to model performance, given that it will also increase the possibility of overfitting and introducing more noise. In addition, it is a common assumption that idiographic models (ie, individualized predictive models) would perform better in capturing each individual's pain rating compared with population-level models. However, it remains unclear how much data are required to capture the within-individual variability to ensure the generalizability of idiographic models. With a mediocre amount of data, it is possible that idiographic models do not outperform population-level models. Lastly, studies have been suggesting larger sample sizes benefit predictive modeling,[23] but researchers may be interested in determining the specific amount of data required to reach a desired level of model performance, considering the tradeoff between costs and outcomes.

2. Materials and Methods

To investigate these unresolved questions, we first conducted a systematic survey of 57 published research articles featuring neuroimaging-based predictive models of pain. We also conducted benchmark analyses on a large-scale functional magnetic resonance imaging (fMRI) pain dataset (N = 124), in which we delivered thermal stimuli to induce heat-induced pain and collected pain intensity ratings. Unlike previous studies that usually examined the impact of each modeling option in isolation, here, we systematically compared them all using a single large-scale pain fMRI dataset.

2.1. Literature survey

We conducted a literature survey to examine the current state of neuroimaging-based pain biomarker research. For more focused analyses, we only included the papers that used MRI or electroencephalogram (EEG) for pain prediction. **Figure 2** shows a flow diagram of the article search and inclusion. We conducted a search using PubMed for research articles published between January 2008 and August 2020 with the following search terms: "pain" in the Title/Abstract; "predict" or "classf*" in Title/Abstract; "eeg," "fmri," "magnetic resonance imaging," or "brain" in Title/Abstract; "machine learning," "predictive modeling," "decoding," "signature," "svm," or "biomarker" in Title/Abstract; NOT "review" in Publication Type; NOT "Symposium" in Title/Abstract. The number of initially searched articles was 137. The exclusion criteria included (1) nonhuman animal studies, (2) absence of prediction models, (3) studies not about pain, (4) nonempirical research articles (eg, review articles), and (5) studies using other imaging modalities, such as positron emission tomography (PET) and functional near-infrared apectroscopy (fNIRS). We included structural MRI because of its more frequent use in pain biomarker studies, while we excluded PET or fNIRS due to the limited number of studies utilizing these imaging modalities. In addition, PET and fNIRS were less relevant to our

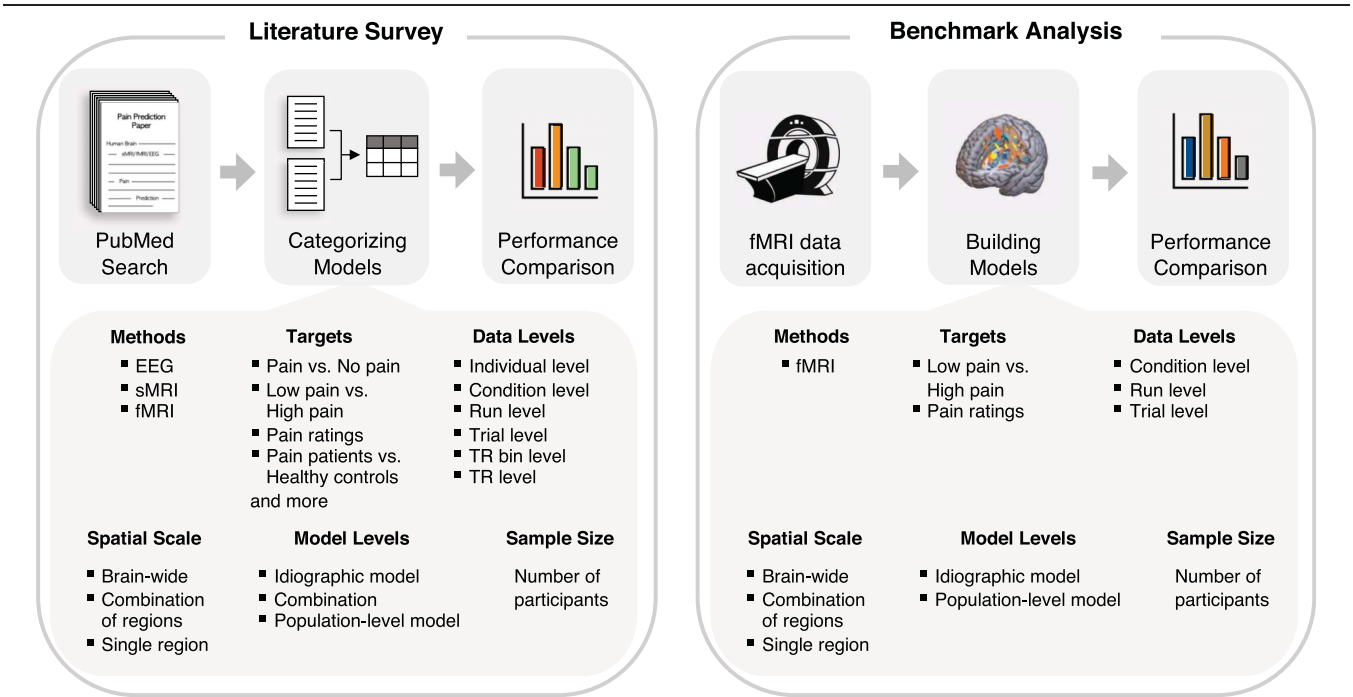

Figure 1. Study overview. This study used 2 approaches to investigate neuroimaging-based pain biomarkers: (1) literature survey and (2) benchmark analysis. For the literature survey, we conducted a literature search on neuroimaging-based biomarkers of pain on PubMed. We then categorized and summarized the predictive models of pain in terms of 12 aspects, including neuroimaging methods, modeling targets, levels of data, spatial scale, levels of the model, and their sample sizes, tasks, algorithms, etc. In this figure, we displayed more details of 6 aspects among them. Each aspect contained multiple categories. We then compared model performances based on the selected aspects. For benchmark analysis, we analyzed a large-scale task-based functional magnetic resonance imaging (fMRI) dataset with painful thermal stimulation. We developed multiple predictive models with varying modeling options that appeared important in the literature survey and evaluated the influences of these modeling options on model performance.

benchmark dataset, which consists of fMRI data. In addition to the article search through PubMed, we manually added 7 research articles that the searched articles cited but were not identified through the PubMed search. You can find the full list of the surveyed papers in Table S2, http://links.lww.com/PAIN/C129. Fifty-seven articles were included for further analyses.

We categorized the predictive models reported in surveyed articles based on the following aspects: (1) measurement tools (EEG, structural MRI, or fMRI), (2) populations (clinical or healthy), and the types of clinical pain, (3) prediction tasks (ie, classification or regression), (4) targets (eg, classification of pain vs no pain, prediction of pain intensity, etc), (5) model levels (eg, idiographic model or population-level model), (6) (train and test) data levels (eg, trial level, run level, etc), (7) spatial scales (eg, single region, brain-wide, etc), (8) experimental tasks (eg, resting state, phasic pain, etc), (9) feature types (eg, activation pattern, connectivity, etc), (10) algorithms (eg, linear support vector machine [SVM], linear regression, etc), (11) validation methods (eg, k-fold cross-validation, leave-one out validation, etc), (12) sample sizes. The full list of the aspects and categories can be found in Table S3, http://links.lww.com/PAIN/C129. For the prediction task, we excluded the multiclass classification task because there was only one model for this category,[38] and thus all models fell into binary classification or regression. For model performance, we chose to compare the classification accuracy and prediction-outcome correlation (ie, a correlation between the predicted and actual values), which were the main performance metric that most of the surveyed models adopted (89.5% and 64.3%, respectively; Figure S1, http://links.lww.com/PAIN/C129). The final comparisons included the 129 training models and 44 independent tests from 57 studies.

After we divided the model performance into 2 categories based on prediction tasks (ie, binary classification or regression), we first compared the performances based on the modeling targets (ie, what the models are designed to predict). Next, given that most of the models (63.6%) focused on the following target categories, including "Pain vs no pain" and "Low pain vs high pain" for binary classification and "Pain rating" for regression, we compared the model performance of these targets for different aspects, including "(train) data level," "spatial scales," "model levels," and "sample size."

2.2. Participants for benchmark analysis

We conducted a benchmark analysis, in which we compared the performance of models with specific modeling options against those with alternative options to evaluate the influence of modeling choices on performance. We utilized a locally collected, large-scale fMRI dataset, which included painful heat stimuli. Multiple models were trained and tested using this dataset, each incorporating a variety of modeling options. We recruited a total of 137 healthy and right-handed participants with no history of neurological, psychiatric, or chronic pain disorders. Among them, 13 participants were excluded due to (1) technical issues (eg, thermal stimulus equipment errors), (2) voluntary discontinuation of the scanning session by participants (eg, intolerable stimulus), or (3) finding of an abnormal structure in the brain (eg, Arachnoid cyst). The final number of participants included in this study was 124 (61 women, age = 22.17 ± 2.69 years [mean ± SD]). This study was approved by the Institutional Review Board at Sungkyunkwan University, and all participants provided written informed consent.

2.3. Experimental design and procedure

We conducted the experiment over 2 visits. On the first visit, participants visited the laboratory to complete a series of self-

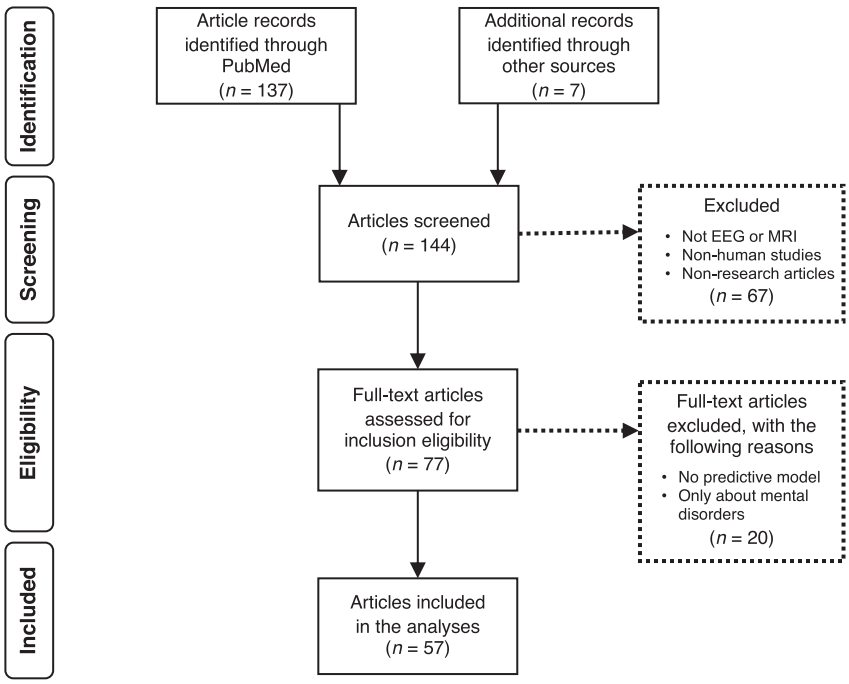

Figure 2. Flowchart of the article selection process for inclusion in this study. This flowchart outlines the systematic approach for selecting articles from initial identification through PubMed and other sources to the final inclusion for model comparison. Research articles from other sources include the articles that were cited by the searched articles but were not identified through the PubMed search. The process includes a comprehensive search, screening of titles and abstracts, eligibility assessment of full-text articles, and the exclusion criteria leading to the final selection. The diagram also shows the numbers of the selected articles for each step. The details of the survey results are shown in Figures 4–6.

report questionnaires. Within 2 weeks, the participants returned to the laboratory, completed another series of self-report questionnaires, and underwent an fMRI experiment. The fMRI experimental procedure consisted of 4 types of runs: (1) resting state, (2) thermal stimulus without movie stimuli, (3) thermal stimulus preceded by a short (20-second) movie clip, and (4) oral capsaicin stimulus. In this study, we used only the runs involving thermal stimulus, ie, (2) and (3) above, and the structural scans.

The thermal stimulation runs (with or without movie stimuli) consisted of 12 trials. In each trial, participants experienced a thermal stimulus for 12 seconds (for details, see "Thermal Stimulation" below) while fixating their eye gaze on a cross shown on the screen. After each thermal stimulus, participants rated the magnitude of their painful experience using the generalized labeled magnitude scale (gLMS).[2] This has a 0 to 1 numerical continuous rating scale with anchors of "Not at all" (0), "A little bit" (0.061), "Moderately" (0.172), "Strongly" (0.354), "Very Strongly" (0.533), and "Most (Strongest imaginable sensation/unpleasantness of any kind)" (1). We removed these anchors and labels during the experiment to prevent the categorical rating behavior.[15] We used the gLMS because of the well-known power function relationship between stimulus and pain intensity,[29,33] which describes the nonlinear increase in pain intensity as stimulus intensity increases. This typically results in a skewed and nonnormal distribution of pain ratings when using linear scales such as visual analog scale and numerical rating scale. The use of gLMS can mitigate the skewness of the distribution of pain ratings by adopting a quasi-logarithmic spacing between labels.[2] In addition, the gLMS allows for better capture of highly sensitive individuals or trials by adding more space at high levels of pain intensity.

Participants completed a total of 8 runs of thermal stimulation, 2 of which were the thermal stimulation runs without a movie and 6 runs with a movie. In the case of the thermal stimulation run with a movie clip, a 20-second movie clip was shown before the thermal stimulation. The inclusion of the movie stimuli was part of a larger study investigating the effects of prestimulus brain states on pain perception, with the movie stimuli used to manipulate the prestimulus brain state. The experimental design, in which the movie clips were presented before the thermal stimulation, allows us to separately analyze the stimulus period and the prestimulus period. In this study, we did not analyze movie-related brain activity. **Figure 3A** shows the trial structure with time information. The movie clips were from a Korean film titled "Summer, Bus" (available at https://youtu.be/-MliIE5PGrI). We split the 12-minute movie into 20-second video clips (ie, 36 movie clips). The sequence of the experimental runs was structured as follows: The initial run was the thermal stimulation run without a movie, followed by 6 consecutive runs of thermal stimulation with a movie clip, and concluding with a final run of thermal stimulation without a movie. Each participant received a total of 96 thermal stimuli (ie, 8 runs and 12 trials for each run). Lastly, for 24 participants, we had to discard one or 2 runs that had some technical issues (eg, sound or thermode malfunction, etc).

2.4. Thermal stimulation

In the thermal stimulation runs, we delivered thermal stimulation using an MRI-compatible PATHWAY Advance Thermal Stimulation system (Medoc Ltd, Ramat Yishay, Israel) with a 16 × 16 mm^2 thermode. We marked 4 different sites on the left forearm of each participant for thermal stimulation. For each run, one of the 4 sites was selected, and the same site was never used in 2 subsequent runs. The order of the stimulation sites was counterbalanced across participants. We delivered thermal stimulation with fixed temperatures ranging from 45°C to 47.5°C in 0.5°C increments. Before the start of each run, we applied the highest temperature (ie, 47.5°C) on a skin site for the run. This was to ensure consistent pain responses throughout the experiment session based on a previous study, where Jepma et al.[16] observed that when a high-intensity stimulus was delivered in the middle of an experimental run for the first time, the pain response before and after the stimulus became qualitatively different. Thus, the delivery of the highest temperature before each run aimed to avoid the site-specific habituation effects in the middle of runs. On each trial, the stimulation was delivered for 12 seconds (2.5 seconds ramp-up, 7 seconds at plateau temperature, 2.5 seconds ramp-down) from the baseline temperature (32°C).

2.5. Functional magnetic resonance imaging data acquisition and preprocessing

The fMRI data were collected using a 3T Siemens Prisma scanner at the Center for Neuroscience Imaging Research, Institute for Basic Science, Sungkyunkwan University. Structural T1-weighted images were obtained using a magnetization-prepared rapid gradient echo sequence (0.7 × 0.7 × 0.7 mm^3 voxel size, repetition time: 2400 ms, echo time: 2.34 ms, slice thickness: 0.70 mm, flip angle: 8°, field of view: 224 × 224 mm^2, inversion time: 1150 ms). Functional data were then acquired using gradient echo-planar imaging sequence (2.7 × 2.7 × 2.7 mm^3 voxel size, repetition time: 460 ms, echo time: 27.20 ms, flip angle: 44° slice thickness: 2.7 mm, slices, field of view: 220 × 220 mm^2, order of slice accession: interleaved). The first 18 image volumes of each run were removed before image preprocessing for image intensity stabilization. Structural and functional MRI data were preprocessed using our in-house preprocessing pipeline (https://github.com/cocoanlab/human-fmri_preproc_bids) based on Statistical Parametric Mapping 12 (SPM12) software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12), FMRIB Software Library (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), and independent component analysis (ICA)-automatic removal of motion artifacts (ICA-based strategy for AROMA) software (https://github.com/maartenmennes/ICA-AROMA). Structural T1-weighted images were coregistered to the single-band reference functional image for each subject using normalized mutual information and then normalized to the Montreal Neurological Institute space using SPM12. For functional echo-planar imaging preprocessing, the pipeline included the following steps: motion correction (realignment), distortion correction using FMRIB Software Library's top-up, spatial normalization to Montreal Neurological Institute space using coregistered T1-weighted images with the interpolation to 2 × 2 × 2 mm^3 voxels, spatial smoothing with a Gaussian kernel (5-mm full-width half-maximum), and ICA to automatically detect and remove participant-specific, motion-related artifacts (ICA-AROMA).[30] In a quality control phase, a few runs were excluded based on the following 2 criteria based on framewise displacement (FD): (1) the average FD of a run exceeds 0.2 mm, and (2) the FD of any volume was greater than 5 mm in a run.[27,28]

2.6. Single-trial analysis

Before conducting the predictive modeling analysis, we estimated single-trial response magnitudes for each voxel using a general linear model (GLM) with separate regressors for each trial, as in the "beta series" approach[31] (**Fig. 3B**). We constructed each trial regressor for movie watching, pain anticipation, and heat

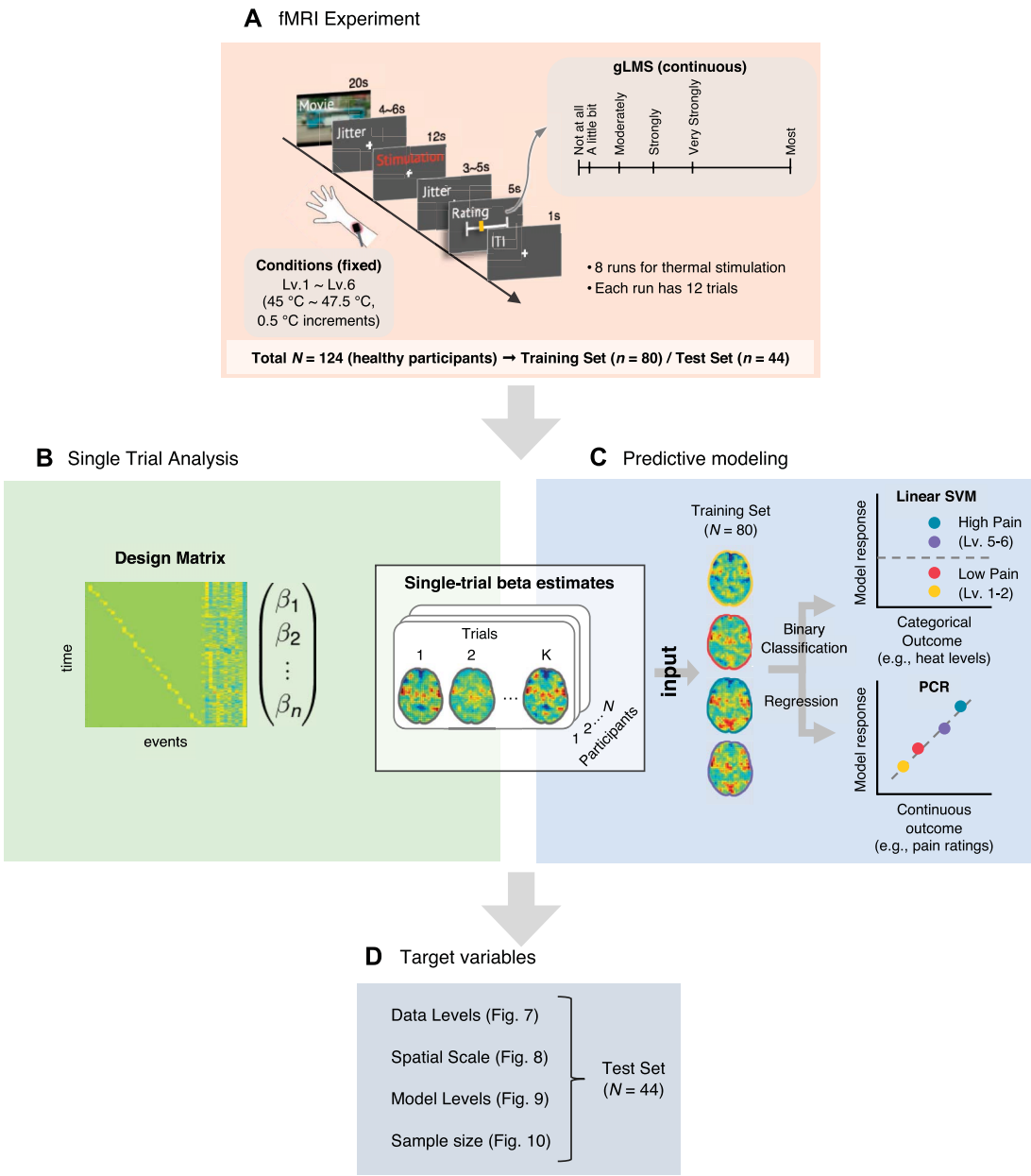

Figure 3. Overview of benchmark analysis. We performed a benchmark analysis to examine the impact of modeling options on model performance. (A) Functional magnetic resonance imaging experiment setup. We collected task-based fMRI data from 124 healthy participants. We delivered thermal stimulation with 6 levels of stimulus intensity, ranging from 45 to 47.5°C in 0.5°C increments. After the heat stimulation, participants were asked to rate pain intensity on the generalized labeled magnitude scale (gLMS).[2] Participants completed a total of 8 runs of thermal stimulation, 2 of which were without a movie and 6 runs with a movie. In the case of runs with a movie, a 20-second movie clip was shown before the thermal stimulation. Each run consisted of 12 trials. For model training and testing, we divided the dataset into a training set (N = 80) and an independent test set (N = 44). All results were based on the independent test dataset. (B) Single-trial analysis. We obtained single-trial voxel-wise beta maps for each participant using a general linear model (GLM) with separate regressors for each trial, as in the "beta series" approach.[31] These beta maps served as inputs for subsequent analyses. (C) Predictive modeling. We develop predictive models with machine learning techniques. For binary classification, the target was "high" vs "low" pain. We defined "high" pain as heat stimulus levels 5 and 6 and "low" pain as heat stimulus levels 1 and 2. For binary classification, we used linear support vector machines (linear SVMs). For regression-based prediction, the target was "pain ratings," and we used principal component regression (PCR) for model training. (D) Target variables. We provide results about 4 aspects (ie, data levels, spatial scales, model levels, and sample size). The details of the results are shown in the following figures (Figs. 7–10). fMRI, functional magnetic resonance imaging.

stimulation with a boxcar convolved with SPM12's canonical hemodynamic response function. We also included one regressor for the pain rating period for each run. In the preprocessing, since we already removed participant-specific, motion-related artifacts in ICA-AROMA, we additionally regressed out only the following nuisance covariates—5 principal components of white matter and cerebrospinal fluid signal and a linear trend. We then calculated trial-by-trial variance inflation factors (VIFs), which measure design-induced uncertainty due to collinearity with nuisance regressors. This step was crucial to identify trials potentially influenced by artifacts. Trials with VIFs exceeding 3 were excluded from further analyses. On average, 0.1371 trials were excluded per participant due to high VIFs, with a standard deviation of 0.7686. The single-trial beta maps served as inputs for predictive modeling.

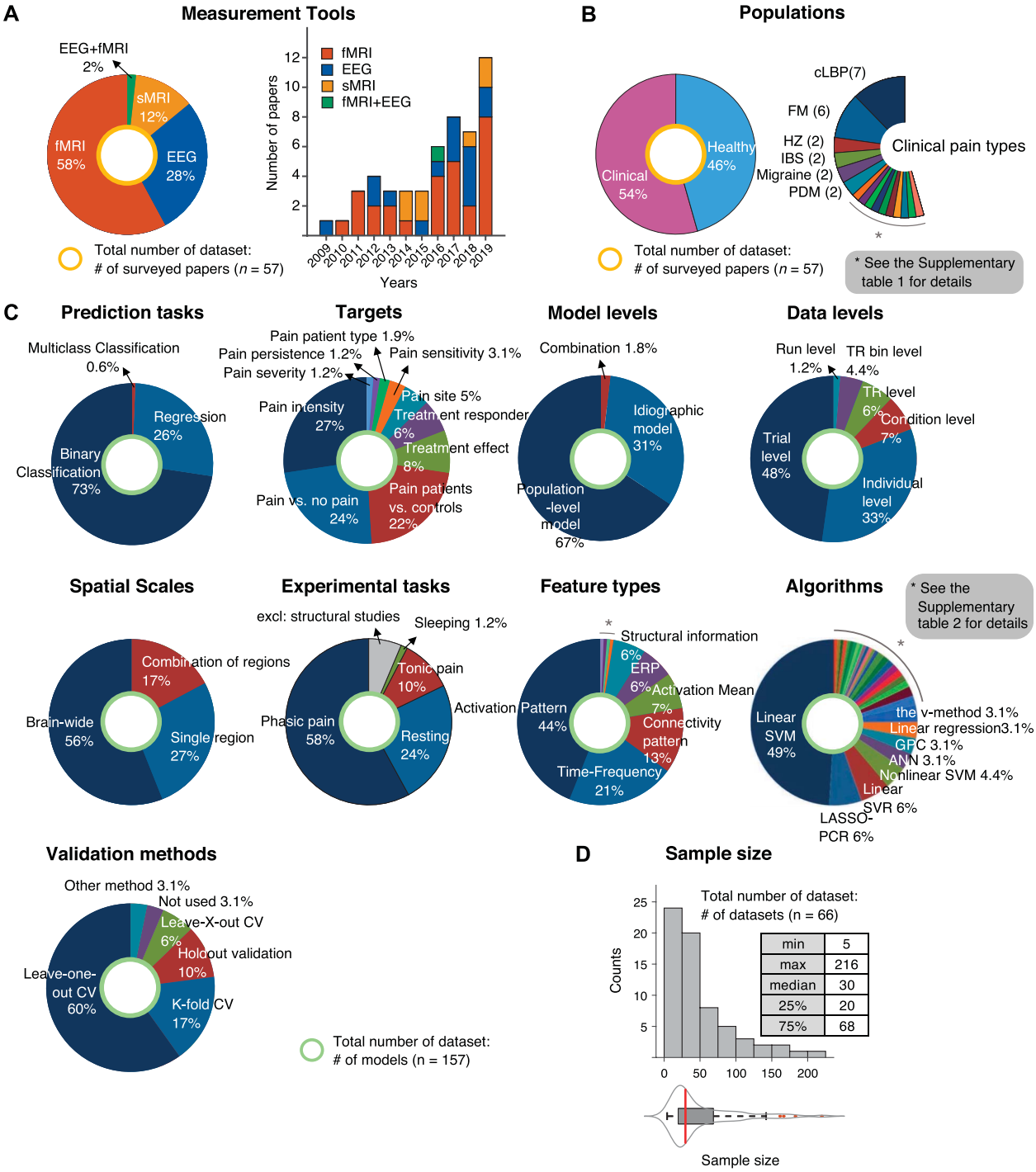

Figure 4. Survey results of neuroimaging-based predictive models of pain. (A) This illustrates the distribution of measurement tools used in the final 57 studies selected for the survey, highlighted by the yellow circle. It also shows the distribution of the publication years of these studies. (B) The pie charts, based on the same set of studies, detail the proportion of healthy vs clinical populations used to train predictive models and categorize the types of clinical pain investigated. (C) Out of the full-text review, 157 models were included for analysis, as indicated by the green circle. The pie charts break down the categories of prediction tasks, model levels, spatial scales, experimental tasks, feature types, validation methods, and algorithms employed. Studies on structural neuroimaging data were omitted from the "Experimental tasks." (D) The box plot represents the sample size distribution across 66 unique datasets from 57 studies. The red line indicates the median value. ANN, artificial neural network; cLBP, chronic low back pain; FM, fibromyalgia; GPC, Gaussian processes classification; HZ, herpes zoster; IBS, irritable bowel syndrome; LASSO-PCR, least absolute shrinkage and selection operator-principal component regression; PDM, primary dysmenorrhea; TMD, temporomandibular disorders; SVM, support vector machine; SVR, support vector regression.

2.7. Predictive modeling

We trained predictive models for benchmark analysis based on single-trial beta images. There were 2 types of tasks: (1) binary classification of high pain vs low pain, and (2) regression of pain ratings (**Fig. 3C**). For binary classification, we defined "high" pain as heat stimulus levels 5 and 6 and "low" pain as heat stimulus levels 1 and 2. We trained classifier models using linear SVMs implemented in "fitcsvm.m" function from the MATLAB Statistics and Machine Learning

Toolbox (https://www.mathworks.com/help/stats/support-vector-machine-classification.html). An SVM classifies data by identifying the optimal hyperplane that separates data points of one class from those of the other class, with distances from the hyperplane indicating the likelihood that an input belongs to one class or the other class.[14] For a soft margin parameter, we used the default value of C = 1. For regression, we trained models to predict pain ratings provided by participants using the gLMS. The principal component regression (PCR) algorithm was employed to estimate pain ratings from single-trial fMRI data. The PCR begins by performing principal components analysis on the input data to reduce its dimensionality and then linearly fits the component scores to the training data. We selected the minimum number of principal components that explained more than 80% of the variance.

We partitioned the data into training (n = 80) and testing datasets (n = 44). To guarantee that the training and testing datasets were comparable in terms of the outcome variable's distribution, we employed a random shuffling for the participants. The final division of datasets was selected based on the criterion of having nonsignificant differences in pain ratings between the datasets, as determined by 2-sample *t* tests. We did not use any of the testing data during the model training, and all testing results were based on the n = 44 hold-out independent test datasets.

2.8. Benchmark analysis

In the benchmark analysis, we selected 4 benchmark scopes that corresponded with the 4 aspects defined in the survey: (1) data level, (2) spatial scale, (3) model level, and (4) sample size.

(1) Data level: The "data level" indicates the number of trials that were averaged for model training and testing. Each participant had a total of 96 single-trial beta estimates and pain ratings if there was no issue, such as technical problems or high VIFs. The 96 trials consisted of 2 repeats of the same stimulus within a run, a total of 8 runs, and 6 stimulus intensity levels (ie, $96 = 2 \times 8 \times 6$). We established 5 data levels by averaging different numbers of trials in the training and testing datasets, respectively. First, "No Average" means that we used a total of 96 single-trial beta images and pain ratings without averaging (ie, trial level). Second, we averaged 2 trials with a stimulus with the same intensity within a run, resulting in 48 brain images and pain ratings per participant (ie, run level). Third, we averaged 4 trials with the same intensity across 2 runs, generating 24 brain images and pain ratings per participant. Fourth, we averaged 8 trials with the same intensity from odd runs (ie, run 1, 3, 5, 7) and even runs (ie, run 2, 4, 6, 8), which yielded 12 brain images and pain ratings per participant. Finally, we averaged 16 trials (ie, 2 repeats within a run and 8 repeats across runs), resulting in 6 brain images and pain ratings per participant (ie, condition level). In addition, we performed an additional analysis with the matched number of testing data (which was 6 data points) for the different test data levels for fair comparisons across different test data levels. For the comparisons of different data levels, we fixed other benchmark scopes—*gray matter* for the spatial scale, *population-level* modeling for the model level, and *full data* (training data n = 80) for sample size.

(2) Spatial scale: To examine the effects of the spatial scale, we used (1) 21 predefined pain-predictive regions of interest (ROIs) obtained from a previous study[17] (Figure S2, http://links.lww.com/PAIN/C129), (2) a meta-analytic map associated with the term "pain" obtained from Neurosynth.org (association test; downloaded on November 1, 2021),[44] and (3) a gray matter mask (GM). In the benchmark analysis for the spatial scale, we tested the impact of increasing spatial scale by randomly selecting one pain-predictive region out of 21 ROIs, then randomly adding more regions incrementally (1, 3, 6, 10, and 15) until all 21 ROIs were included for model training and testing. We repeated this process 100 times and obtained the mean accuracy for each spatial scale and for each participant. For the comparisons of different spatial scales, we fixed other benchmark scopes—*run level* for the data level, *population-level* modeling for the model level, and *full data* (training data n = 80) for sample size.

(3) Model level: We compared idiographic vs population-level predictive models. In this analysis, the sample size of the training dataset was reduced to n = 61 because we needed the data to have the complete 8-run data for the analysis described below. For idiographic modeling (ie, within-individual prediction modeling), we trained a model based on 6-run data and tested the model on the remaining 2-run data. For the data split, we also used the 2-sample *t* test to ensure that there was no significant difference in the outcome variable (ie, pain ratings) between the training and testing datasets. We also trained one population-level model based on 6-run data concatenated across all 61 participants. We tested these 2 types of models on 2 different types of testing datasets. One was the remaining 2 hold-out run data from 61 participants as described above, and the other was all run data from 44 hold-out participants. To test the idiographic models on the 44 hold-out participants' data, we averaged the idiographic models to construct one predictive model. For the comparisons of the results, we fixed other benchmark scopes—*trial level* for the data level, *gray matter* for the spatial scale, and n = 61 for sample size.

(4) Sample size: Lastly, we evaluated the impact of varying sample sizes using a random selection procedure akin to that employed for evaluating spatial scale. Specifically, we examined the impact of increasing sample sizes by randomly selecting 10 participants from the total pool of 80 participants, then incrementally adding more participants in steps (10, 20, 30, …, 70) until all 80 participants were included in the model training and testing. We repeated this process 100 times and obtained the mean accuracy for each sample size and for each iteration. For the comparisons of different sample sizes, we fixed other benchmark scopes—*run level* for the data level, *population-level* for the model level, and *gray matter* for the spatial scale.

3. Results

3.1. Survey results on the use of modeling options

Figure 2 shows the article selection process, and **Figures 4–6** show the literature survey results. First, as shown in **Figure 4A**, the fMRI was the most popular neuroimaging modality for modeling—ie, 57.9% out of the 57 studies used fMRI. The number of research on neuroimaging-based pain biomarkers has increased since the first publication in 2009. Note that the plot does not display the number of publications of 2020 (n = 7) and 2021 (n = 1) given that our survey covers only 8 months of 2020 (our survey period was between January 2008 and August 2020). **Figure 4B** shows the study populations of the 57 studies. A greater number of studies addressed clinical pain outcomes (54.4%) as compared with healthy ones (45.6%), and chronic low back pain was the most popular among clinical pain conditions (7 studies). **Figure 4C** shows the survey results of 157 predictive

models from 57 studies. The survey highlights several key trends in the field: (1) Prediction task: Binary classification models (72.6%) were predominant, outnumbering regression models (26.7%) by approximately 3-fold. (2) Modeling targets: Over half of the models focused on "pain intensity" (27.4%) and "pain vs no pain" (23.5%). "pain patients vs controls" was also an important target (21.6%). Note that "pain intensity" included the following 2 subtargets, "pain ratings" and "high pain vs low pain," which were used for further benchmark analysis. (3) Model level: A majority of the models were population level (66.8%), which was almost double the number of idiographic models (31.2%). This result suggests that studies on pain biomarkers usually focus on their clinical applications. (4) Data level: Trial level (47.7%) and individual level (33.1%) were most commonly used. (5) Spatial scale: Brain-wide models were more prevalent (56.0%) compared with region-based models, including single-region (26.7%) and multiregion models (17.2%). (6) Experimental task: The phasic pain task was the most common (57.9%), followed by resting-state tasks (24.2%). (7) Feature type: Activation patterns were the most frequently used features (43.9%). (8) Algorithm: Linear SVM was the most utilized algorithm (49.0%), followed by PCR with Lasso regularization (6.3%) and linear support vector regression (5.7%). Notably, the top 3 algorithms were all linear methods. (9) Validation methods: Leave-one-out cross-validation was most common (59.8%), with K-fold cross-validation next (17.2%). **Figure 4D** depicts the distribution of sample sizes across 66 unique datasets. The median sample size was 30, indicative of the typical scale for fMRI experiments.

3.2. Survey results on model performance

We then compared the model performances across different modeling targets and options. For these comparisons, we focused on the following 6 aspects—prediction tasks, modeling targets (**Fig. 5**), data and model levels, spatial scale, and sample size (**Fig. 6**). These aspects are among the key aspects that can provide references for future studies and are also included in the later benchmark analysis.

First, we compared the classification accuracy and prediction-outcome correlation (the correlation between predicted and actual values) across various modeling targets. To present these results effectively, the plots in **Figure 5** are organized by the maximum performance values achieved by the training models for each specific target. The most frequently studied targets were "pain patients vs controls" (26 training models and 14 independent tests), "pain vs no pain" (35 training models and 12 independent tests), and "pain rating prediction" (18 training models and 9 independent tests). "High pain vs low pain" was also a popular target, but there was no independent test (13 training models and no independent test). "Treatment responder," "treatment effect," and "pain site" were the next popular targets, but these also had one or no independent test, highlighting that these targets need further validation studies. For "pain patients vs controls" and "pain rating prediction," the model performances for the independent tests (median accuracy = 67.3%, median $r = 0.32$) were lower than the training results (accuracy = 72.4%, $r = 0.725$), indicating the potential bias in the reported model performances. For "pain vs no pain," some independent tests showed higher performances than training, which was from one specific study that showed high accuracy in multiple independent tests.[41] Note that targets with a large number of reports showed high variance in their model performance, suggesting that the targets with a small number of reports may not be able to serve as references for future studies and require further studies. In addition, the model performance of "pain rating prediction" exhibited a bimodal distribution in training results (**Fig. 5B**). This, along with the high variance in the classification model performances, implies that other variables and modeling options may have significant impacts on model performance.

We then evaluated the influences of modeling options on the model performance, as shown in **Figure 6**. Here are some observations. First, the survey on the train data levels showed that the condition level results were higher than the trial-level results (median accuracy = 87.3% for the condition level training models and 69.0% for the trial-level training models; **Fig. 6A**). Second, also from **Figure 6A**, while classification models did not exhibit significant decreases in performance in independent tests, regression models showed an overall decline in performance, suggesting that regression models may have more difficulty in generalizing than classification models. Third, for the spatial scale, as shown in **Figure 6B**, the brain-wide models showed higher performance (median accuracy = 74.8% for brain-wide training models) than a combination of regions (71.6%) or single region (63.7%). Fourth, for the model levels (**Fig. 6C**), the population-level models showed the highest performance both in classification and regression models (median accuracy = 74.5% and median $r = 0.74$ for population-level training models). Last, there was a significant negative relationship between sample size and model performance of regression models, $r = -0.8227$, $P = 0.000\ 000\ 1$ (**Fig. 6D**). This negative relationship was not observed in classification models, $r = 0.2913$, $P = 0.0239$.

3.3. Benchmark analysis

Given that the different experimental designs and populations across surveyed studies limited the direct comparability of the results, we additionally performed benchmark analyses on the locally collected large-scale pain fMRI dataset (n = 124) with a single experimental design. Our benchmark analysis focused on the following 4 aspects: (1) data level, (2) spatial scale, (3) model level, and (4) sample size, and 2 different targets: binary classification of high vs low pain and the prediction of pain ratings with regression models. **Figure 3** summarizes the benchmark analysis pipeline. Importantly, all the results presented in this study were obtained from tests on the hold-out test set of 44 participants.

3.3.1. Benchmark analysis (1): data level

The first benchmark analysis was on the data level. The "data level" indicates the number of trials averaged for model training and testing (**Figs. 7A and B**). **Figure 7C** provides the classification model performance (ie, accuracy), and **Figure 7D** provides the regression model performance (ie, prediction-outcome correlation). In **Figure 7C and D**, the plots in the first and second columns present the benchmark analysis results for the train vs test data levels, respectively, focusing on the impact of different train vs test data levels on model performance. The plots in the third column show the benchmark analysis results for the test data levels but with the matched number of test data, which was 6 data points.

For the high vs low pain classification, we found that model performance decreased as more data were averaged in the training dataset (**Fig. 7C**). For example, the model performances across 5 different train data levels were 85.05% ± 1.31% (mean accuracy ± SEM) for "no average," 84.53% ± 1.32% for 2 trials

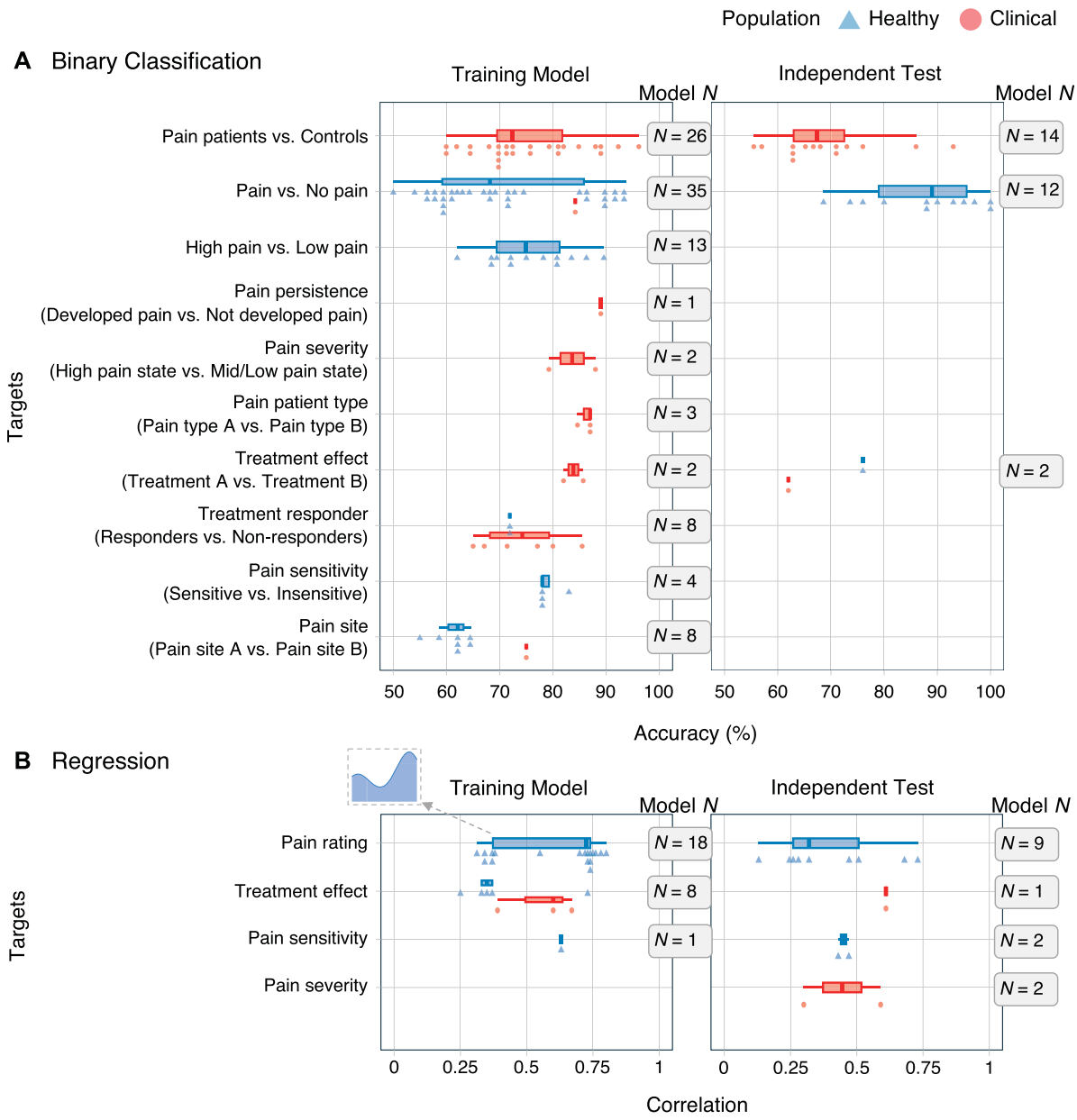

Figure 5. Survey results 1: Model performance across different modeling targets. (A) Binary classification task. The box plots illustrate performance distributions for 10 different modeling targets during model training and independent testing phases. The colors of box plots and dots denote healthy (blue) and clinical (red) populations. Dots represent each model's classification accuracy (%), with triangles for healthy populations and circles for clinical populations. "Model *N*" indicates the number of models included in the comparisons. The targets were ordered by their peak performance values. Note that there have been studies on healthy populations focusing on the targets of treatment effect and responder. For the "treatment effect" target, an independent test aimed to distinguish between high and low expectancy.[5] For the "treatment responder" target, a study trained models to classify "drug responder" and "placebo responder" in healthy populations.[12] (B) Regression task. Prediction performances of regression-based models were presented for 4 regression targets. We used prediction-outcome correlation (ie, a correlation between the predicted and actual values) for comparison. For the "treatment effect" target, there have been studies on healthy populations aimed at predicting placebo analgesia.[40] A histogram for "Pain rating" in the training model set is provided, highlighting the bimodal distribution of model performance for the target.

averaged, 83.17% ± 1.43% for 4 trials averaged, 81.70% ± 1.44% for 8 trials averaged, and 80.31% ± 1.45% for 16 trials averaged. The same patterns were also observed in the other test data level (ie, 16 trials averaged). A multilevel GLM showed that the trend of decreasing model performance over averaging the training data was significant, $\beta = -1.09$, $z = -3.53$, $P = 0.000\,42$, 2-tailed, bootstrap test. The same pattern was also observed when 16 trials were averaged for the testing data, $\beta = -5.09$, $z = -4.50$, $P = 0.000\,01$, multilevel GLM, 2-tailed, bootstrap test.

By contrast, we observed that the model performances increased as more data were averaged in the testing dataset. For example, when the train data level was fixed at "no average," the model performances across 5 different test data levels were 85.05% ± 1.31% (mean accuracy ± SEM) for "no average," 88.36% ± 1.52% for 2 trials averaged, 91.57% ± 1.62% for 4 trials averaged, 92.90% ± 2.05% for 8 trials averaged, and 93.75% ± 2.17% for 16 trials averaged. A multilevel GLM showed that the trend of increasing model performance over averaging the testing data was significant, $\beta = 2.35$, $z = 3.43$, $P = 0.000\,61$, 2-tailed, bootstrap test. The same pattern was observed for 16 trials averaged for the training data, $\beta = 2.28$, $z = 3.40$, $P = 0.000\,68$, multilevel GLM, 2-tailed, bootstrap test. Furthermore, the

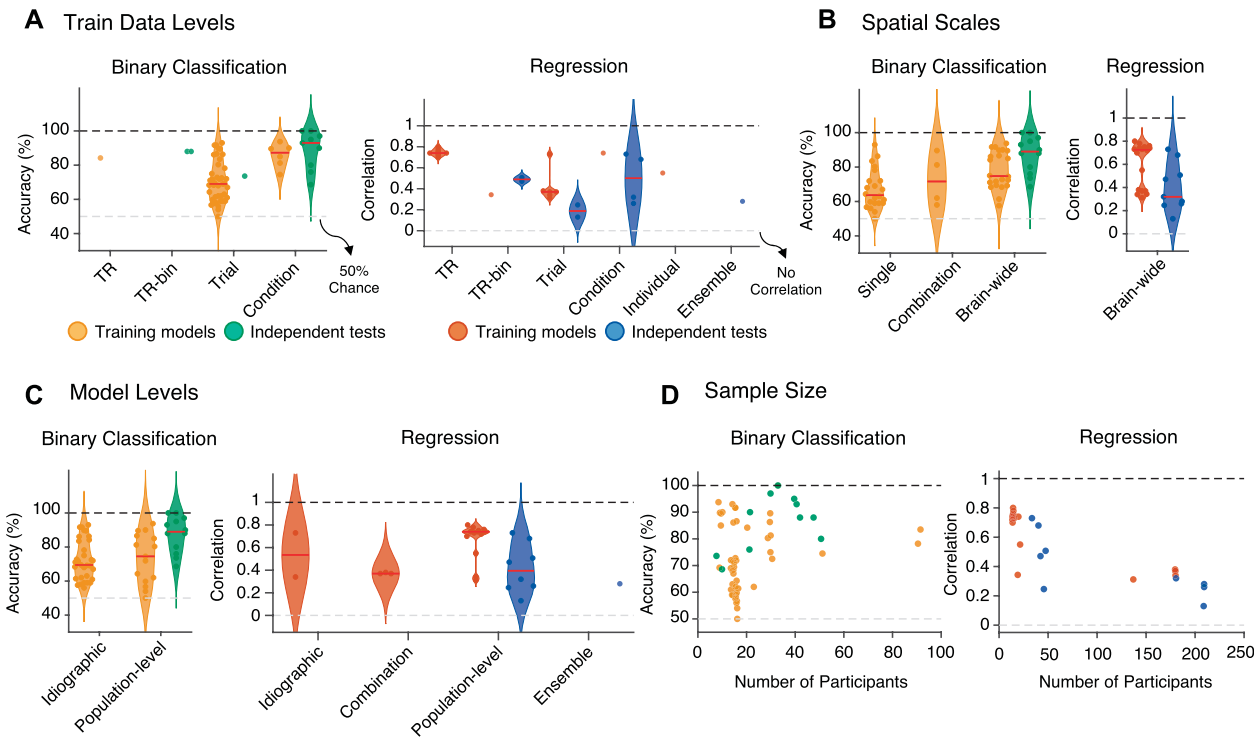

Figure 6. Survey results 2: model performance across different data levels, spatial scales, model levels, and sample sizes. (A–C) Violin plots illustrate the distribution of model performance across various aspects: training data levels (A), spatial scales (B), and model levels (C). In these plots, individual dots represent the performance of each model, quantified as accuracy for binary classification tasks or prediction-outcome correlation for regression tasks. Yellow and red plots represent the performance derived from the training datasets, while green and blue plots illustrate the performance observed in independent tests. The median performance for each category is indicated by red lines within the plot. A black dashed line represents the possible maximum performance, and a gray dashed line denotes the baseline level of performance, which corresponds to chance or no correlation. (D) Impact of sample size on model performance. Scatter plots demonstrate how the number of participants in a model influences performance, with separate visualizations for binary classification and regression tasks. Yellow and red dots represent the performance derived from the training datasets, while green and blue dots denote the performance in independent tests.

same trend was observed when we compared the performance across 5 different test data levels with a matched number of test data points (ie, 6 data points per participant) for both "no average" and 16 trials averaged for the training data, βs = 4.56 and 3.49, *z*s = 3.78 and 3.18, *P*s = 0.000 15, and 0.001 47, multilevel GLM, 2-tailed, bootstrap test.

Lastly, the main findings for the pain rating regression were similar to the binary classification results—the model prediction performance increased as more data were averaged in the test dataset, while it decreased as more data were averaged in the training dataset. These effects were significant and consistent across different combinations of train and test data levels (**Fig. 7D**). For example, multilevel GLMs with a linear regressor revealed significant increases in model performance across different test data levels for all train data levels—"no average" (β = 0.07, *z* = 4.43, *P* = 0.000 01, 2-tailed, bootstrap) and "16 trials averaged" (β = 0.09, *z* = 3.93, *P* = 0.000 09). The details of the model performance and significant test results are in Table S4, http://links.lww.com/PAIN/C129.

3.3.2. Benchmark analysis (2): spatial scale

The second benchmark analysis was on the spatial scale. **Figure 8A** presents an overview of the analysis. While increasing spatial scales, we developed new models and calculated the classification accuracy (**Fig. 8B**) and prediction-outcome correlation (**Fig. 8C**) based on the test dataset (n = 44). The model development was repeated 100 times using random combinations of brain regions. For "brain-wide masks," we employed 3 distinct masks: one encompassing 21 pain-predictive regions ("21") from a previous study,[17] the Neurosynth "Pain" mask (NP), and a GM without iteration. **Figure 8B and C** shows model performance across 44 participants in the independent dataset.

The results indicated that the model performance increased with the increasing numbers of combined regions. For the increasing numbers of combined regions (ie, 1, 3, 6, 10, and 15 regions), the mean binary classification accuracies were 68.79%, 73.01%, 78.67%, 82.56%, and 85.27% (**Fig. 8B** and Table S5, http://links.lww.com/PAIN/C129), and the prediction-outcome correlations were 0.46, 0.57, 0.62, 0.67, and 0.70 (**Fig. 8C** and Table S5, http://links.lww.com/PAIN/C129). These increasing trends of model performance were statistically significant, βs = 4.25 and 0.06, *z*s = 3.66 and 3.58, *P*s = 0.000 25 and 0.000 34, 2-tailed, bootstrap test, multilevel GLM. For the tests for "brain-wide masks," there were no significant differences among the type of brain masks for the binary classification task, 87.31%, 86.96%, and 88.27% for "21," NP, and GM, respectively, all |*t*|s < 0.98, *P*s > 0.33, 2-tailed, paired *t* test. However, there was a significant difference between GM vs "21" for the rating prediction task, prediction-outcome *r* = 0.75 (GM) and 0.71 ("21"), *t*(43) = 2.626, *P* = 0.0119, 2-tailed, paired *t* test, suggesting that the information distributed across the gray matter, even outside of the 21 pain-predictive regions, was helpful for predicting pain ratings. Last, there were no significant differences between "21" vs NP and GM vs NP for the rating prediction task, prediction-outcome *r* for NP = 0.72, *t*(43) = 0.764, *P* = 0.4489 for "21" vs NP, *t*(43) = 1.962, *P* = 0.0562 for GM vs NP.

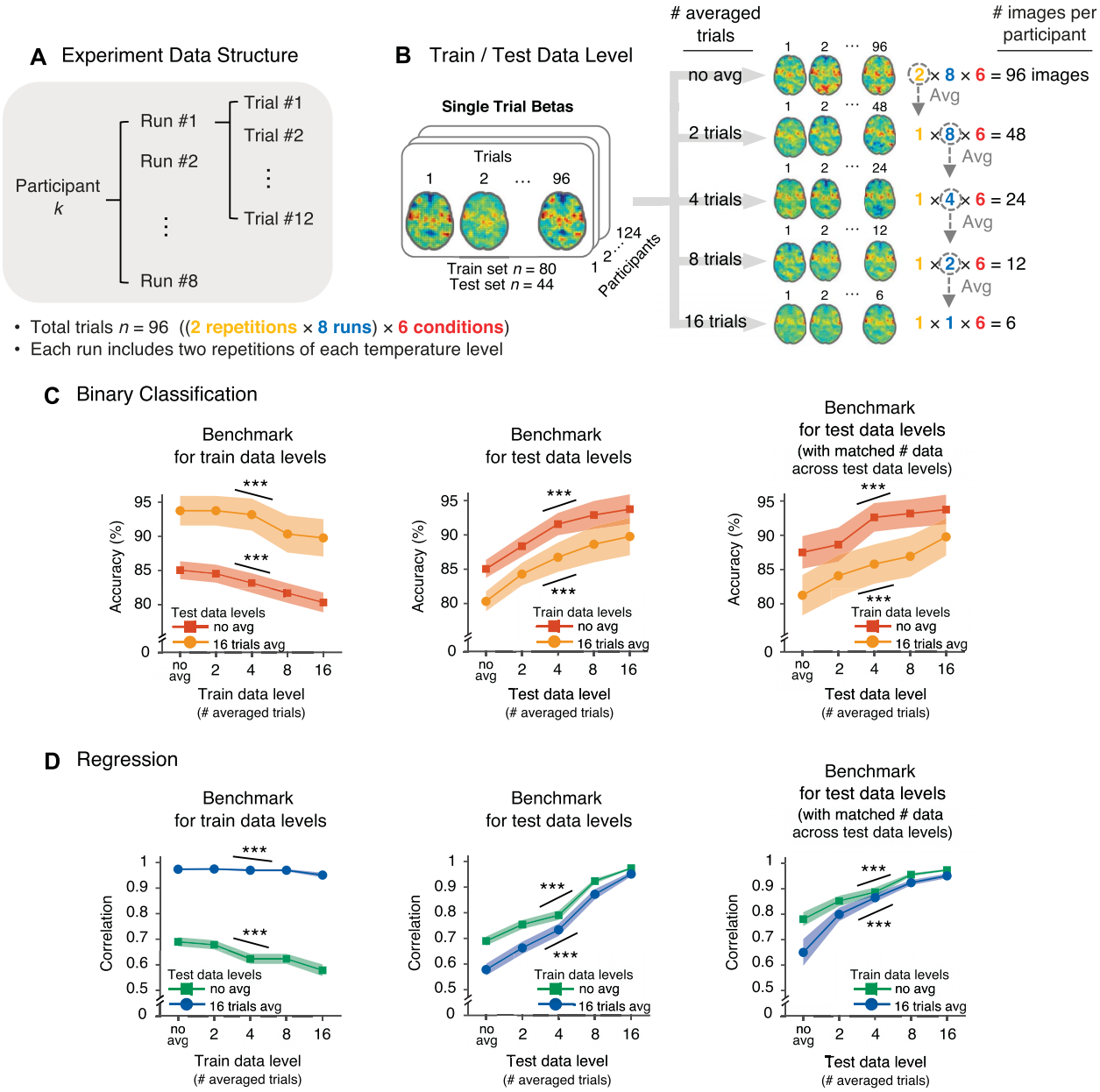

Figure 7. Benchmark analysis results for data levels (ie, data averaging). (A) Data structure. Each participant's data consisted of 96 trials, distributed over 8 runs, with each run comprising 12 trials across 6 temperature conditions, with each temperature level repeated twice. (B) Data were analyzed at 5 averaging levels: single trials (no averaging), and averages of 2, 4, 8, and 16 trials. This averaging was applied separately to both training and testing datasets. (C and D) The plots show the impact of averaging on model performance in the binary classification (C) or the regression (D) tasks. Asterisks indicate the significance of the slopes from the multilevel general linear model (GLM) models assessing the effects of averaging on model performance. Left: The effect of training data averaging on model performance at 2 test data levels ("no averaging" and "16 trials averaged"), Center: The effect of test data averaging on model performance at 2 training data levels ("no averaging" and "16 trials averaged"), Right: The effect of test data averaging on model performance at 2 training data levels, with a fixed number of images in the test set (6 data points). The lines indicate mean accuracy or correlation (respectively) across 44 participants, with shadings representing the standard error of the mean. The color scheme differentiates the averaging conditions with red/green indicating no averaging and yellow/blue indicating 16 trials averaged; we selected these 2 data levels as they were the most commonly used based on our literature survey (ie, single-trial and condition level). Model performance exhibits significant trends with respect to data averaging, as established by a multilevel GLM analysis. Detailed performance metrics and statistical results are provided in Table S4 (http://links.lww.com/PAIN/C129). ***$P < 0.001$.

3.3.3. Benchmark analysis (3): model level

The third benchmark analysis was on the model level. The "model level" analyses compared the test results between the idiographic vs population-level predictive modeling (**Fig. 9A**, for details, see Materials and Methods). Briefly, idiographic predictive modeling indicates the modeling based on a single participant's data, whereas population-level predictive modeling indicates the modeling using the data combined across all individuals. There were also 2 types of independent tests in this benchmark analysis. First, we tested the models on the within-individual hold-out data of the training dataset (n = 61). For this, we held out 2-run data for the testing by using only 6-run data for training. Second, we tested the models on the fully independent test dataset (n = 44). We combined the different training and testing methods, which resulted in the following 4 combinations: ① idiographic model ("Id") tested on within-individual hold-out data, ② population-level model ("Po") tested on within-individual hold-

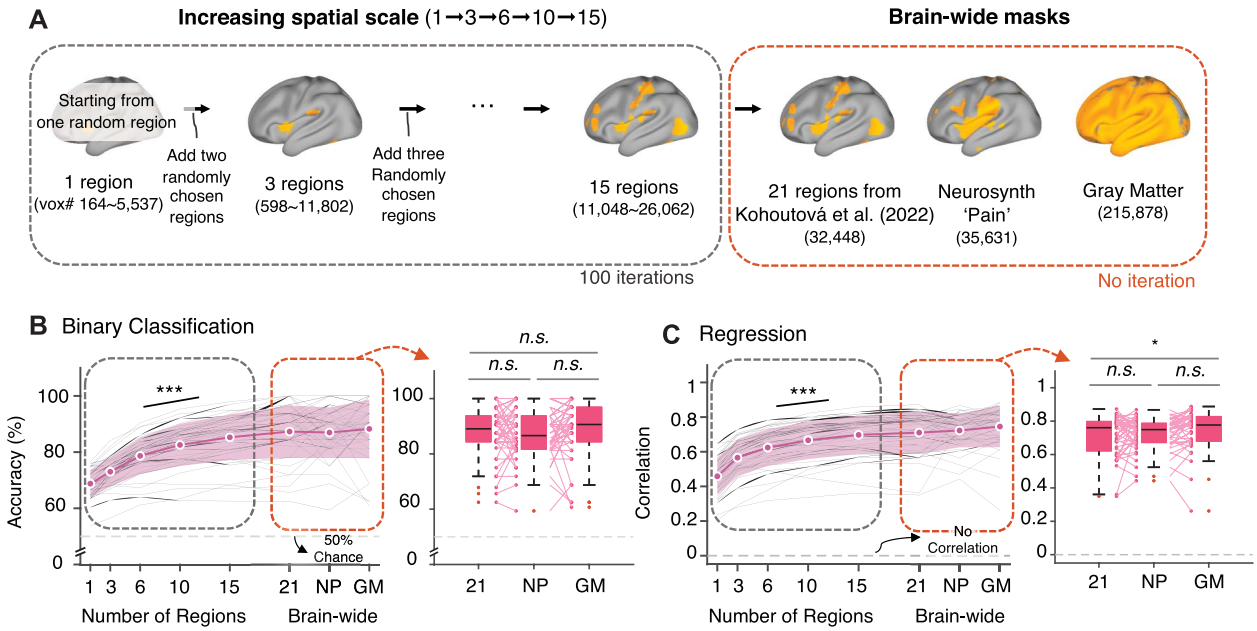

Figure 8. Benchmark analysis results for spatial scales. (A) The analysis employed 2 different approaches. In the first approach, we used a progressive mask construction method. Starting with one randomly selected region, we incrementally added more regions, up to a total of 15. These regions were selected from predefined regions of interest (ROIs) identified in a previous study.[17] Predictive models were then trained using these incrementally constructed masks. This iterative process was repeated 100 times, as indicated in the illustration by a gray dashed line box. The second approach involved the use of 3 comprehensive brain-wide masks. These included: (1) a mask encompassing all 21 ROIs from a previous study,[17] (2) a "Pain" mask derived from Neurosynth (labeled as NP), and (3) a gray matter mask (GM). Each mask's voxel count is noted below its respective label. For the cumulative regions mask, a range is provided instead of a fixed number, reflecting the variability due to the random selection process. This approach is delineated in the box with an orange dashed line. (B and C) The average performance from 100 iterations was calculated for each spatial scale. The reddish-purple lines and shaded areas depict the mean and standard deviation of performance, respectively, across 44 participants in the test dataset. Gray thin lines show the model performance for each participant. Performance trends showed a significant increase with the number of regions, validated by multilevel general linear models (GLMs) with *z*s = 3.66 and 3.58, *P*s = 0.000 25 and 0.000 34 for binary classification (B) and regression (C), respectively, 2-tailed, bootstrap tests. The boxplots show the differences in model performance among the 3 brain-wide masks. n.s. $P > 0.05$, $^{*}P < 0.05$, $^{***}P < 0.001$. GM, Gray matter; NP, Neurosynth pain mask.

out data, ③ averaged idiographic model ("Avg Id") tested on the independent test dataset, and ④ population-level model ("Po") tested on the independent test dataset. **Figure 9B and C** shows the comparisons of the classification and the regression model performances. We conducted 2-sample *t* tests to compare model performance between the idiographic vs population-level models.

In the binary classification, the performance of idiographic and population-level models did not significantly differ, $t(120) = 0.271$, $P = 0.7868$ for testing on the within-individual hold-out data, $t(86) = 1.194$, $P = 0.2358$ for testing on the independent test dataset, 2-tailed, 2-sample *t* test. In the regression analysis, while the idiographic and population-level models demonstrated comparable performance on within-individual hold-out data, $t(120) = 1.287$, $P = 0.2005$, the population-level model significantly outperformed the idiographic model on the independent test dataset, $t(86) = 2.813$, $P = 0.0061$, 2-tailed, 2-sample *t* test. Table S6, http://links.lww.com/PAIN/C129 provides detailed results of model performance.

3.3.4. Benchmark analysis (4): sample size

The final benchmark analysis focused on the impact of sample size (**Fig. 10A**). While increasing the sample size, we developed new models and calculated the classification accuracy (**Fig. 10B**) and prediction-outcome correlation (**Fig. 10C**) based on the independent test dataset (n = 44). The model development was repeated 100 times using a random selection of participants. ("Increasing sample size" in **Fig. 10A**). For "full sample size," all 80 participants from the training dataset were used without the random selection (and thus no iteration). The model performances on the independent test dataset (n = 44) are shown in **Figure 10B and C**.

The test results for the "increasing sample size" condition showed that the model performance increased with the increasing number of training samples. The binary classification accuracies (mean ± SD) for the increasing numbers of combined samples (ie, 10, 20, …, 60, and 70 participants) were 83.50% ± 0.58, 85.39% ± 1.09, 85.94% ± 0.83, 86.86% ± 0.70, 87.15% ± 0.40, 87.59% ± 0.34, 87.99% ± 0.40 (**Fig. 10B** and Table S7, http://links.lww.com/PAIN/C129), and the prediction-outcome correlations were 0.698 ± 0.01, 0.727 ± 0.01, 0.738 ± 0.01, 0.741 ± 0.01, 0.744 ± 0.01, 0.745 ± 0.01, and 0.746 ± 0.004 (**Fig. 10C** and Table S7, http://links.lww.com/PAIN/C129). These increasing trends of model performance were statistically significant (βs = 0.68 and 0.01, *z*s = 3.96 and 4.02, *P*s = 0.000 08 and 0.000 06, 2-tailed, bootstrap test, multilevel GLM).

4. Discussion

In this study, we conducted a literature survey and 4 benchmark analyses on neuroimaging-based pain biomarkers. The importance of building pain biomarkers lies in the need for objective assessment of pain because of its subjective nature of pain assessment, which hampers the choice of appropriate interventions or the development of new treatments. A recent study conducted a qualitative review on the topic,[37] but more systematic analyses on the choices of modeling targets and options have yet to be done. To fill the gap, we systematically

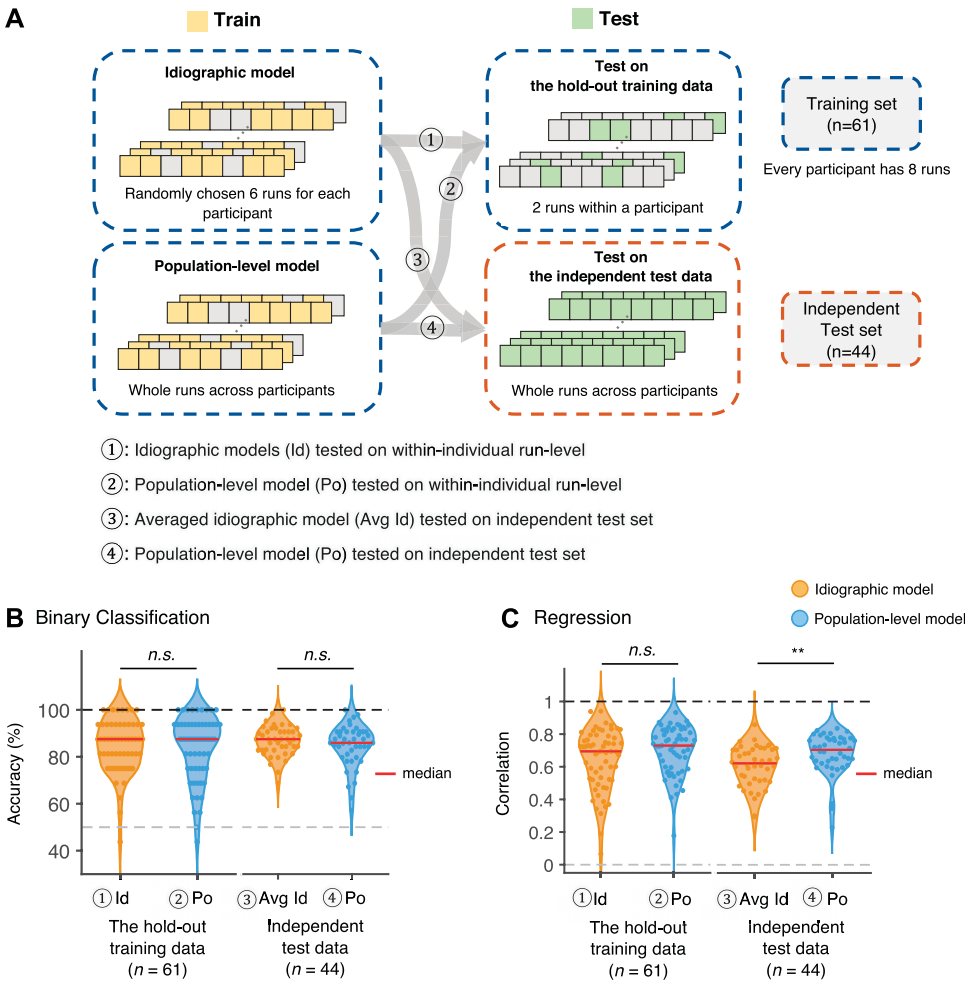

Figure 9. Benchmark analysis results for model levels. We provide a benchmark analysis of model performance at the individual (idiographic) and group (population) levels. In this analysis, we used data from 6 runs for training and data from 2 runs for testing. This choice required the use of data from only 61 participants who had complete data for all 8 runs. (A) Model training and testing procedures. Idiographic model (Id): Predictive models were trained on trial-level data from 6 runs per participant and tested on unseen data from the remaining 2 runs (①). In addition, we created a group-level model by averaging all individual models and tested the model on an independent test dataset (③). Population-level model (Po): Models were developed using trial-level data from all participants in the training dataset. Again, the data from 6 runs were used. The model was tested on both unseen data from the remaining 2 runs of 61 participants (②) and data from an independent test dataset (④). (B) Binary classification. No significant performance differences were found between the idiographic and population-level models when evaluated on both the hold-out training data and the independent test dataset (*P*-values = 0.7868 for ① vs ② and 0.2358 for ③ vs ④, 2-sample *t* test). (C) Regression. There was no significant difference in performance on the hold-out training data ($P = 0.2005$ for ① vs ②); however, the population-level model significantly outperformed the averaged idiographic model on the independent test dataset, $P = 0.0061$ for ③ vs ④, 2-sampled *t* test. $^{**}P < 0.01$.

compared the performance of different models from previously published studies on neuroimaging-based pain biomarkers and conducted benchmark analyses using a large-scale fMRI pain dataset. In these analyses, we focused on the following modeling aspects—prediction tasks, modeling targets, data levels, spatial scales, model levels, and sample sizes. The primary goal of this study was to provide a bird's eye view of neuroimaging-based pain prediction and a useful guide for making decisions about modeling options.

Our survey results provide multiple interesting observations. First, the survey revealed a few preferred targets and options for predictive modeling. The majority of the predictive models were designed for classification (73%), aimed at population-level prediction (66%), used brain-wide features (56%), and were trained on the data at the trial level or individual level (80%). These characteristics suggest that those who developed the models were mindful of their clinical applications and utility[10] because population-level diagnostic models that can be applied to new individuals are favorable for clinical purposes. Second, fair comparisons of model performance across different modeling targets were quite challenging due to the small number of studies for each target. Even with numerous studies conducted on the modeling target, the model performance remained highly variable, suggesting that many factors influenced the level of model performance. Thus, it is not simple to determine the level of difficulty for different modeling targets. In addition, consideration should be given to the FDA-NIH Biomarker Working Group's Biomarkers, EndpointS, and other Tools categories based on their intended use. It is likely that the levels of difficulty are varied across these biomarker categories, and thus, the standards for clinical translation should be carefully determined for different targets. The third observation from the survey pertains to the limited number of independent tests. Only a small proportion of studies (28%) assessed their models on independent datasets, making it challenging to assess the impacts of modeling targets and options on model performances in an unbiased way. Last, we

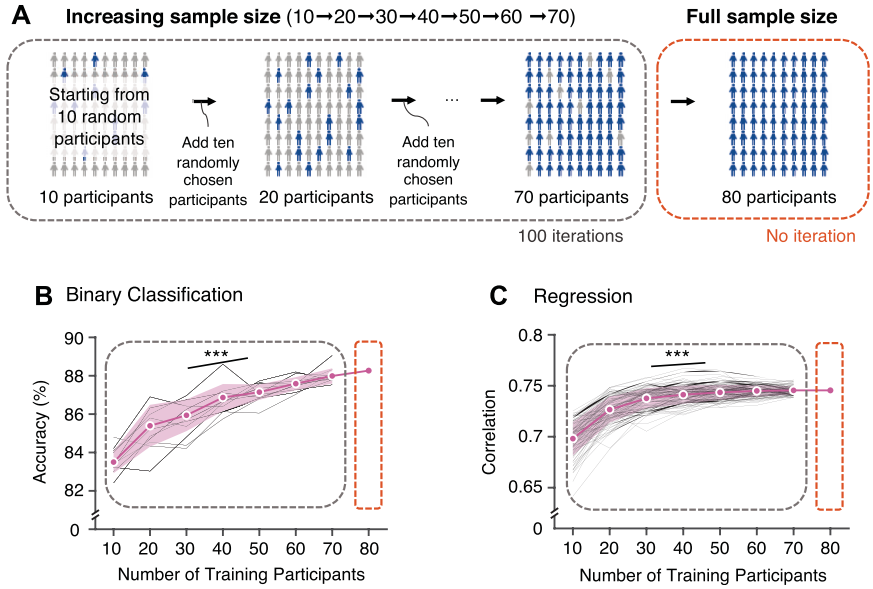

Figure 10. Benchmark analysis results for sample size. We investigated how the number of participants in the training set affects the performance of predictive models. Models were iteratively developed with increasing training sample sizes, ranging from 10 to 70 participants, and compared with a model trained with the full set of 80 participants. (A) This study began with 10 randomly selected participants from the training dataset and incrementally and randomly added sets of 10 participants to build new models. This incremental process was repeated 100 times, as indicated by the box with a gray dashed line. In a separate analysis, the full cohort of 80 participants was used, as indicated by the box with an orange dashed line. (B and C) Average model performance across iterations is shown by the reddish-purple line, with the shaded area indicating the standard deviation, for the binary classification (B) and regression (C) tasks. Gray thin lines show model performance in each iteration. Statistically significant improvements in model performance were observed as the number of training participants increased (for the binary classification and regression tasks, zs = 3.96 and 4.02, Ps = 0.000 08 and 0.000 06, respectively, 2-tailed, bootstrap test). The full sample size models performed better than the other models with smaller sample sizes, achieving 88.26% accuracy for binary classification and a correlation of 0.74 for regression. ***$P < 0.001$.

observed the negative relationship between sample size and model performance in the regression models. This negative relationship could be interpreted as a well-known phenomenon called "funnel plot asymmetry"[32] or "winner's curse,"[6] which suggests potential publication bias in the literature. However, the interpretation requires caution, given that such negative relationships were not observed in the classification models or for other modeling options, such as spatial scales.

Our benchmark analysis results also offer several interesting observations regarding the effects of modeling options on model performance. First, the distinct effects of data averaging on model performance were observed for training vs testing data—less data averaging was helpful for training data, whereas more data averaging was helpful for testing data. This can be understood as a tradeoff between increasing the signal-to-noise ratio by averaging vs expanding the sample distribution. That is, expanding the sample distribution of training data improves model performance, but reducing the noise of testing data also improves model performance. Second, both survey and benchmark analysis results failed to show the benefits of idiographic modeling compared with population-level models. This is somewhat contradictory to the common notion that personalized models should perform better than population models because the idiographic approach removes one important source of variance—between-individual variability.[22] Considering the recent suggestions about the extensive sampling of small N participants,[13,24] the nonsignificant results may be due to the small amount of data per individual in our sample. Thus, future studies should examine the effects of idiographic vs population-level modeling with multisession data from a small number of participants. Third, our results showed that including more brain regions and increasing sample sizes improve model performance, consistent with previous studies.[4,18,23] However, in the benchmark analysis, the results suggested that including extra brain regions beyond the regions known to be important for pain prediction did not always guarantee higher model performance. For example, although the number of voxels within the GM (n_{vox} = 215,878) was almost 7 times greater than 21 a priori pain-predictive regions (n_{vox} = 32,448) or Neurosynth mask (n_{vox} = 35,631), the classification performance of the GM-based model was not significantly better than the models based on the 21 pain-predictive regions and Neurosynth mask. However, when it comes to the regression task, the model based on the GM showed significantly better performance than the 21 region-based regression model, suggesting that the characteristics of the prediction task can affect the benefit of including more regions and voxels.

There are multiple limitations in this study. First, our survey results were limited by the small number of studies. For example, only 28% of the surveyed studies reported test results with independent test datasets, which made it difficult to obtain unbiased estimates of model performances. This means that our survey results could be biased. In addition, owing to the insufficient number of studies in the survey, we could not conduct fair comparisons between results from healthy and clinical populations despite the substantial differences in their brain conditions.[1] Therefore, future studies with larger sample sizes and a more balanced representation of healthy and clinical populations are needed to draw more definitive conclusions. In addition, given that brain connectivity and its dynamics have been implicated in clinical pain,[19,20,36] the generalizability of our benchmark analyses, which were based on activation patterns,

to clinical pain also requires further investigation. Second, some studies with more than one predictive model could have a greater influence on the survey results. When we compared model performance, we compiled the testing results at the model level but noted that some studies provided multiple models while others provided only a single model. Thus, to take this data structure into account, a hierarchical approach to performance comparisons should be considered in future studies. Table S2, http://links.lww.com/PAIN/C129 provides the number of models and independent tests for each study. Third, we had to use the most popular performance measures, such as correlation, to compare model performances, but it is well known that correlation as a performance measure has some caveats (eg, being insensitive to scaling, providing biased results, etc.).[26] As more researchers adopt better practices in predictive modeling, we should be able to use better performance measures, such as R^2, for the comparisons. Fourth, in the benchmark analyses, we did not cover all the aspects considered in the survey due to the limitations of our dataset. For example, our dataset only included fMRI data from healthy participants and used thermal stimulation to induce pain, and thus, we could not provide results on other neuroimaging or stimulus modalities or clinical samples. Considering that the inclusion of multimodal data to develop composite biomarker signatures is an important and promising future direction,[34] our analysis focusing only on fMRI and thermal pain data is certainly limited. In addition, our dataset is based on experimental pain, where pain intensity is strongly influenced by stimulus intensity. Thus, caution is warranted when generalizing our results to other neuroimaging modalities or clinical pain contexts where pain is dissociated from the stimulus. Finally, our survey includes articles published between 2008 and 2020. Given the recent rise in popularity of machine learning and artificial intelligence, there may be many recent publications that we have missed. Thus, future studies are needed to update our findings.

In conclusion, this study investigated the influences of modeling options and targets on the performance of neuroimaging-based pain biomarkers. Through a literature survey and benchmark analyses, we found that data levels, spatial scales, and sample sizes were important determinants of classification and prediction performance. To improve model performance, incorporating a larger number of pain-related brain regions, increasing the sample sizes, and reducing data averaging in the training dataset while increasing it in the test dataset appeared to be helpful. These findings will serve as a useful reference for making decisions on neuroimaging-based biomarker development, highlighting the importance of a careful selection of modeling variables to build better-performing neuroimaging pain biomarkers. Furthermore, our findings would offer useful insights for their potential translation into clinical settings by presenting a bird's eye view of the field of neuroimaging-based pain prediction.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Data availability: All data used to generate main figures are available at https://doi.org/10.5281/zenodo.10432300.

Code availability: The codes for generating the main figures are available at https://doi.org/10.5281/zenodo.10432300. In-house MATLAB codes for fMRI data analyses are available at https://github.com/canlab/CanlabCore and https://github.com/cocoanlab/cocoanCORE.

Acknowledgments

This work was supported by IBS-R015-D1 (Institute for Basic Science; to C.-W.W.) and 2021M3E5D2A01022515 (National Research Foundation of Korea; to C.-W.W.).

Author contributions: S.L. and C.-W.W. designed the experiment. D.H.L. collected the data. D.H.L. and S.L. preprocessed the data. D.H.L. and C.-W.W. analyzed the data, interpreted the results, and wrote the manuscript. C.-W.W. edited the manuscript and provided the supervision.

Supplemental digital content

Supplemental digital content associated with this article can be found online at http://links.lww.com/PAIN/C129.

Article history:
Received 12 February 2024
Received in revised form 10 July 2024
Accepted 10 July 2024
Available online 25 September 2024

References

[1] Apkarian AV, Bushnell MC, Treede RD, Zubieta JK. Human brain mechanisms of pain perception and regulation in health and disease. Eur J Pain 2005;9:463–84.
[2] Bartoshuk LM, Duffy VB, Green BG, Hoffman HJ, Ko CW, Lucchina LA, Marks LE, Snyder DJ, Weiffenbach JM. Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. Physiol Behav 2004;82:109–14.
[3] Boly M, Faymonville ME, Schnakers C, Peigneux P, Lambermont B, Phillips C, Lancellotti P, Luxen A, Lamy M, Moonen G, Maquet P, Laureys S. Perception of pain in the minimally conscious state with PET activation: an observational study. Lancet Neurol 2008;7:1013–20.
[4] Brodersen KH, Wiech K, Lomakina EI, Lin CS, Buhmann JM, Bingel U, Ploner M, Stephan KE, Tracey I. Decoding the perception of pain from fMRI using multivariate pattern analysis. Neuroimage 2012;63:1162–70.
[5] Brown CA, Almarzouki AF, Brown RJ, Jones AKP. Neural representations of aversive value encoding in pain catastrophizers. Neuroimage 2019;184:508–19.
[6] Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14:365–76.
[7] Coghill RC, Sang CN, Maisog JM, Iadarola MJ. Pain intensity processing within the human brain: a bilateral, distributed mechanism. J Neurophysiol 1999;82:1934–43.
[8] Coghill RC. The distributed nociceptive system: a framework for understanding pain. Trends Neurosci 2020;43:780–94.
[9] Davis KD, Flor H, Greely HT, Iannetti GD, Mackey S, Ploner M, Pustilnik A, Tracey I, Treede RD, Wager TD. Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. Nat Rev Neurol 2017;13:624–38.
[10] Davis KD, Aghaeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, Burczynski ME, Crean C, Edwards R, Gaudilliere B, Hergenroeder GW, Iadarola MJ, Iyengar S, Jiang Y, Kong JT, Mackey S, Saab CY, Sang CN, Scholz J, Segerdahl M, Tracey I, Veasley C, Wang J, Wager TD, Wasan AD, Pelleymounter MA. Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. Nat Rev Neurol 2020;16:381–400.
[11] FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Bethesda; Silver Spring: Food and Drug Administration (US); National Institutes of Health (US), 2016.
[12] Gram M, Graversen C, Olesen AE, Drewes AM. Machine learning on encephalographic activity may predict opioid analgesia. Eur J Pain 2015;19:1552–61.
[13] Gratton C, Nelson SM, Gordon EM. Brain-behavior correlations: two paths toward reliability. Neuron 2022;110:1446–9.
[14] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. New York: Springer, 2009.

[15] Hayes JE, Allen AL, Bennett SM. Direct comparison of the Generalized Visual Analog Scale (gVAS) and general Labeled Magnitude Scale (gLMS). Food Qual Prefer 2013;28:36–44.
[16] Jepma M, Jones M, Wager TD. The dynamics of pain: evidence for simultaneous site-specific habituation and site-nonspecific sensitization in thermal pain. J Pain 2014;15:734–46.
[17] Kohoutova L, Atlas LY, Buchel C, Buhle JT, Geuter S, Jepma M, Koban L, Krishnan A, Lee DH, Lee S, Roy M, Schafer SM, Schmidt L, Wager TD, Woo CW. Individual variability in brain representations of pain. Nat Neurosci 2022;25:749–59.
[18] Kragel PA, Koban L, Barrett LF, Wager TD. Representation, pattern information, and brain signatures: from neurons to neuroimaging. Neuron 2018;99:257–73.
[19] Kucyi A, Davis KD. The dynamic pain connectome. Trends Neurosci 2015;38:86–95.
[20] Lee JJ, Kim HJ, Ceko M, Park BY, Lee SA, Park H, Roy M, Kim SG, Wager TD, Woo CW. A neuroimaging biomarker for sustained experimental and clinical pain. Nat Med 2021;27:174–82.
[21] Levine FM, Lee De Simone L. The effects of experimenter gender on pain report in male and female subjects. PAIN 1991;44:69–72.
[22] Lindquist MA, Krishnan A, Lopez-Sola M, Jepma M, Woo CW, Koban L, Roy M, Atlas LY, Schmidt L, Chang LJ, Reynolds Losin EA, Eisenbarth H, Ashar YK, Delk E, Wager TD. Group-regularized individual prediction: theory and application to pain. Neuroimage 2017;145:274–87.
[23] Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Hendrickson TJ, Malone SM, Kandala S, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, Conan GM, Uriarte J, Snider K, Lynch BJ, Wilgenbusch JC, Pengo T, Tam A, Chen J, Newbold DJ, Zheng A, Seider NA, Van AN, Metoki A, Chauvin RJ, Laumann TO, Greene DJ, Petersen SE, Garavan H, Thompson WK, Nichols TE, Yeo BTT, Barch DM, Luna B, Fair DA, Dosenbach NUF. Reproducible brain-wide association studies require thousands of individuals. Nature 2022; 603:654–60.
[24] Naselaris T, Allen E, Kay K. Extensive sampling for complete models of individual brains. Curr Opin Behav Sci 2021;40:45–51.
[25] Petre B, Kragel P, Atlas LY, Geuter S, Jepma M, Koban L, Krishnan A, Lopez-Sola M, Losin EAR, Roy M, Woo CW, Wager TD. A multistudy analysis reveals that evoked pain intensity representation is distributed across brain systems. PLoS Biol 2022;20:e3001620.
[26] Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. JAMA Psychiatry 2020;77:534–40.
[27] Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 2012;59:2142–54.
[28] Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 2014;84:320–41.
[29] Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. PAIN 1994;56:217–26.
[30] Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 2015;112:267–77.
[31] Rissman J, Gazzaley A, D'Esposito M. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 2004;23:752–63.
[32] Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, Carpenter J, Rucker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D, Higgins JP. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002.
[33] Stevens SS. On the psychophysical law. Psychol Rev 1957;64:153–81.
[34] Tracey I, Woolf CJ, Andrews NA. Composite pain biomarker signatures for objective assessment and effective treatment. Neuron 2019;101:783–800.
[35] Tracey I. Neuroimaging mechanisms in pain: from discovery to translation. PAIN 2017;158(suppl 1):S115–22.
[36] Vachon-Presseau E, Berger SE, Abdullah TB, Griffith JW, Schnitzer TJ, Apkarian AV. Identification of traits and functional connectivity-based neurotraits of chronic pain. PLoS Biol 2019;17:e3000349.
[37] van der Miesen MM, Lindquist MA, Wager TD. Neuroimaging-based biomarkers for pain: state of the field and current directions. Pain Rep 2019;4:e751.
[38] Vijayakumar V, Case M, Shirinpour S, He B. Quantifying and characterizing tonic thermal pain across subjects from EEG data using random forest models. IEEE Trans Biomed Eng 2017;64:2988–96.
[39] Wager TD, Atlas LY. The neuroscience of placebo effects: connecting context, learning and health. Nat Rev Neurosci 2015;16:403–18.
[40] Wager TD, Atlas LY, Leotti LA, Rilling JK. Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. J Neurosci 2011;31:439–52.
[41] 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: 1388–97.
[42] Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. J Clin Nurs 2005;14:798–804.
[43] Woo CW, Schmidt L, Krishnan A, Jepma M, Roy M, Lindquist MA, Atlas LY, Wager TD. Quantifying cerebral contributions to pain beyond nociception. Nat Commun 2017;8:14211.
[44] Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 2011;8:665–70.