



# Predicting placebo analgesia in patients with chronic pain using natural language processing: a preliminary validation study

Paulo Branco<sup>a,b</sup>, Sara Berger<sup>c,d</sup>, Taha Abdullah<sup>a,b</sup>, Etienne Vachon-Pressseau<sup>e,f</sup>, Guillermo Cecchi<sup>d</sup>, A. Vania Apkarian<sup>a,b,\*</sup>

## Abstract

Patients with chronic pain show large placebo effects in clinical trials, and inert pills can lead to clinically meaningful analgesia that can last from days to weeks. Whether the placebo response can be predicted reliably, and how to best predict it, is still unknown. We have shown previously that placebo responders can be identified through the language content of patients because they speak about their life, and their pain, after a placebo treatment. In this study, we examine whether these language properties are present before placebo treatment and are thus predictive of placebo response and whether a placebo prediction model can also dissociate between placebo and drug responders. We report the fine-tuning of a language model built based on a longitudinal treatment study where patients with chronic back pain received a placebo (study 1) and its validation on an independent study where patients received a placebo or drug (study 2). A model built on language features from an exit interview from study 1 was able to predict, a priori, the placebo response of patients in study 2 (area under the curve = 0.71). Furthermore, the model predicted as placebo responders exhibited an average of 30% pain relief from an inert pill, compared with 3% for those predicted as nonresponders. The model was not able to predict who responded to naproxen nor spontaneous recovery in a no-treatment arm, suggesting specificity of the prediction to placebo. Taken together, our initial findings suggest that placebo response is predictable using ecological and quick measures such as language use.

**Keywords:** Placebo, Chronic pain, Language, Latent semantic analysis, Machine learning, Psycholinguistics

## 1. Introduction

Placebo analgesia is overwhelmingly and increasingly observed in clinical trials for drug development,<sup>14,21,35,39</sup> and in the context of chronic pain, it can lead to long-lasting and clinically significant analgesia<sup>7,8,19,23</sup>—sometimes with the same effect size of drugs specifically indicated for pain relief.<sup>40</sup> Meta-analyses clearly show brain responses to the placebo effect, supporting the idea that placebo is a physiological response that is observable, quantifiable,

and reproducible<sup>44</sup> and driven by biological, contextual, and affective cues.<sup>2</sup> Yet, some subjects respond to placebo, but others do not,<sup>3,25,38</sup> and predicting the placebo analgesic response is not trivial. Although there are several studies “predicting” the placebo response (eg, Refs. 9, 10, 34, see Ref. 18 for a review), there is a scarcity of studies that actually test prediction in unbiased validation or replication studies. Some efforts dedicated to studying stable traits predicting placebo response (including brain and personality measures) have shown decent group-level effects<sup>34,37</sup> but sub-standard predictability at the individual level.<sup>36,38,41</sup> Arguably, the placebo response is driven not only by stable, trait-like individual characteristics such as personality and brain properties<sup>38,41</sup> but also by the patients’ context including treatment expectations and previous experiences,<sup>2,18</sup> and hence, a closer look into these aspects might improve the predictability of placebo response.

The quantitative analysis of patients with chronic pain discourse—ie, how they speak about their pain and their previous medical experiences—is a novel way to probe environmental and psychological factors in an ecological and comprehensive way<sup>4,33</sup> lending utility to placebo prediction. We have recently shown that natural language processing can be used to characterize and quantify the language profile of patients with chronic back pain (CBP) as they speak about themselves, their pain, and their medical experiences, and notably, we were able to identify with 79% accuracy who had received substantial analgesia to a placebo pill vs who did not.<sup>4</sup> However, these results were validated within the same sample (ie, cross-validated) and tested in an exit interview (ie, after the treatment) and thus lacked evidence for generalizability and predictive ability.

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

<sup>a</sup> Department of Neuroscience, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, <sup>b</sup> Center for Translational Pain Research, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, <sup>c</sup> Responsible and Inclusive Technology (Exploratory Sciences Division), IBM Research, Yorktown Heights, NY, United States, <sup>d</sup> Computational Psychiatry and Digital Health (Impact Science Division), IBM Research, Yorktown Heights, NY, United States, <sup>e</sup> Faculty of Dentistry and Department of Anesthesia, McGill University, Montréal, QC, Canada, <sup>f</sup> Alan Edwards Center for Research on Pain (AECRP), McGill University, Montréal, QC, Canada

\*Corresponding author. Address: Department of Neuroscience, Northwestern University Feinberg School of Medicine, 303 E Chicago Ave, Chicago, 60611, IL, United States. Tel.: +1 312-503-0404. E-mail address: a-apkarian@northwestern.edu (A.V. Apkarian).

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://www.painjournalonline.com)).

PAIN 164 (2023) 1078–1086

© 2022 International Association for the Study of Pain

<http://dx.doi.org/10.1097/j.pain.0000000000002808>

In this study, we build on previous evidence and assess the validity and generalizability of a predictive language model to identify placebo responders before treatment. To do so, we first reanalyzed data from our initial placebo prediction study<sup>4</sup> to build a single predictive model using language features. The generalizability of this model was then assessed in a new independent study where the language interview was performed before treatment commencement. We further assess specificity by examining whether a language model can not only predict placebo response but also predict analgesic response to an active treatment of naproxen.

We hypothesized that the same language features identified in our previous work<sup>14</sup> will be able to dissociate between placebo responders and nonresponders and validate in an independent sample. We further hypothesized that the placebo prediction model will be able to predict drug response, but to a lesser degree, given that a drug response has, inherently, a placebo effect associated with it.<sup>36</sup>

## 2. Methods

This study report data from 2 experiments that were part of a randomized control trial (RCT) investigating placebo in patients with chronic low back pain (ClinicalTrials.gov registration ID: NCT02013427). Details of this trial were published elsewhere<sup>36,38</sup> and, for brevity, will be described here summarily: the RCT consisted of 2 independent studies, both included patients with CBP, with initial pain of at least 5 of 10 on a visual analog scale, history of CBP for at least 6 months before study commencement, and no evidence of comorbid pain, neurological, or psychological disorders. All participants stopped concomitant pain medication during the studies. Both experiments were double-blinded, and neither the research staff nor the patient knew what treatment the patients received. The first study was designed to generate and test predictive multimodal models of placebo response, and the second study was designed to validate these results in an independent sample or study. Although parts of study 1 have been previously published,<sup>4</sup> here we report, for the first time, the tuning of a classification model based on quantitative language features from the first study (study 1) and its ability to predict placebo responses in the second study (study 2). Both studies were approved by Northwestern's Institutional Review Board, and all participants signed a consent form. Data from the first published study were used to identify initial targets for this validation study.<sup>4</sup>

### 2.1. Participants and study design

#### 2.1.1. Study 1 participants and trial design

The first study assessed the eligibility of 129 participants with CBP. One hundred twenty-five patients with CBP enrolled in this study, and 66 patients completed all aspects of this study including the language interview. After enrolling in this study, subjects rated their pain twice daily for 2 weeks to establish baseline pain using a smartphone app and were then randomized into a no-treatment or treatment arm. The no-treatment group ( $n = 20$ ) was used to control for spontaneous recovery and regression to the mean. The treatment group ( $N = 46$ ) consisted of an active treatment group (naproxen, 500 mg + esomeprazole, 20 mg) and a placebo group (2 lactose pills). Pills were visually identical to ensure blinding of the subject and research staff. Most participants were assigned to the placebo arm ( $N = 42$ ) given that the goal of this study was to study the placebo response; the

active treatment ( $N = 4$ ) was used purely as a blinding tool, and these data were not analyzed. Subjects completed 2 two-week treatment periods followed by a 1-week washout (Fig. 1A). By the end of this study, subjects were interviewed to extract language parameters (exit interview, see below). Demographic and clinical data for study 1 sample are presented in the supplementary material (available at <http://links.lww.com/PAIN/B734>). Further details about this study can be consulted in Ref. 4.

#### 2.1.2. Study 2 participants and trial design

In the second study, 181 patients were assessed for eligibility, 94 patients enrolled in this study, and 50 patients completed all aspects of this study including the language interview. Again, after enrolling in this study, subjects were asked to rate their pain twice daily for 2 weeks and were then randomized to a no-treatment arm or 1 of 2 treatment groups. The no-treatment group here consisted of 5 subjects. The treatment groups ( $N = 42$ , Fig. 1B) consisted of an active treatment group (naproxen, 500 mg + esomeprazole, 20 mg,  $N = 22$ ) and a placebo group (lactose pills,  $N = 20$ ). Demographic and clinical data for the study 2 sample are presented in the supplementary material (available at <http://links.lww.com/PAIN/B734>). Unlike study 1, study 2 subjects were randomized to drug and placebo arms in equal proportions to assess the specificity of the model, that is, if a placebo-based predictive model could also predict drug response and if the drug response and the placebo response were additive. Subjects completed 1 treatment period of 6 weeks while rating their pain twice a day. The language interview here was conducted at the beginning of this study (and unlike study 1 is thus not affected by treatment response). This was performed intentionally so that our model can predetermine placebo response before any treatment, thus being predictive *stricto sensu*.

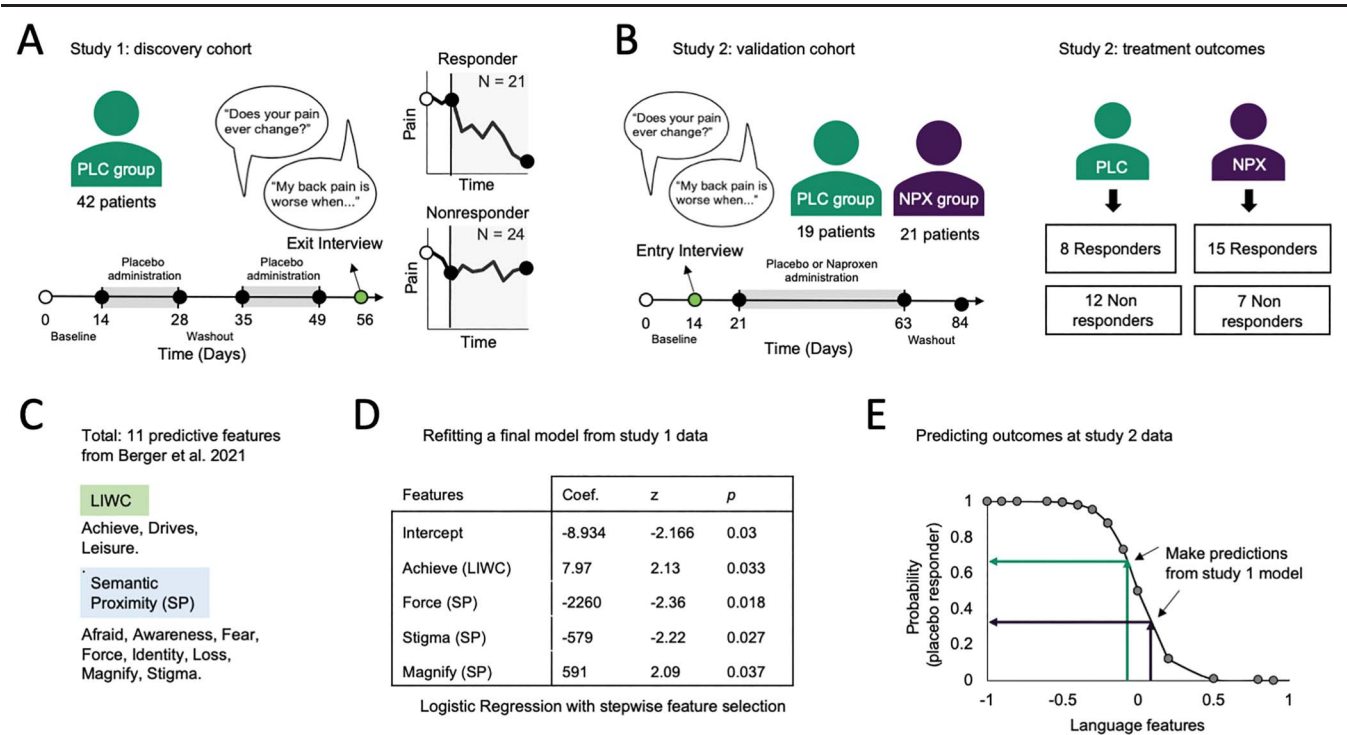
### 2.2. Language interview design and implementation

#### 2.2.1. Study 1 interview

The interview included a warm-up section (3 questions with generic questions) and a main section that probed participants about their pain, emotions, and medical experiences (13 questions). Supplementary Figure 1 (available at <http://links.lww.com/PAIN/B734>) shows the interview script. The interviews lasted  $27.2 \pm 10.3$  minutes. Interviews were semistructured and open ended, to allow conversation to flow as naturally as possible. More details about the interview of study 1 can be seen in detail in Ref. 4.

#### 2.2.2. Study 2 interview

Based on the first study results, the interview was reduced to contain only questions that were deemed important.<sup>4</sup> Thus, in the study 2 interview, we asked only 4 questions to the subjects: "Please describe yourself"; "Please describe a recent event you took part in, recently"; "Please describe your pain"; and "Please describe your previous experiences in the medical system." This was also performed because of logistical reasons, given the limited time we had available with the patients before treatment commencement. Although this adds a nontrivial change in the protocol, we argue that it favors generalizability, that is, if the findings from study 1 replicate even with a substantially smaller interview asking only a subset of questions, it further solidifies the overall robustness of the predictive model. The average duration



**Figure 1.** Study design, model building, and validation. (A) On a first study, 42 patients with chronic back pain received placebo treatment for 2 weeks, followed by a 1-week washout period and a second placebo treatment. One week after the end of the second treatment period, patients were interviewed. Based on the differences in pain between treatment and baseline periods, patients were labelled as placebo responders (N = 21) and nonresponders (24). (B) A second study was performed to validate the study 1 model. In this study, 42 patients with chronic back pain were randomly assigned to a placebo and a drug (naproxen) group. Subjects enrolled in this study rated their baseline pain for 2 weeks and then were interviewed before receiving 2 weeks of placebo or drug treatment. Forty-three percent of patients responded to placebo, and 74% responded to naproxen. (C) Based on our previous study,<sup>4</sup> we selected 11 language features that predicted placebo response. Three from LIWC and 6 from semantic proximity metrics. (D) A new model was generated by performing a bidirectional stepwise linear regression on 11 language features selected a priori based on previous findings. Four features were selected after stepwise elimination, resulting in 91% classification accuracy. (E) The logistic model derived from study 1 was used to predict the probability of placebo response in study 2, using the model described at panel (D). LIWC, Linguistic Inquiry Word Count; NPX, Naproxen; PLC, placebo.

of the interview was significantly shorter ( $3.27 \pm 1.67$  minutes) but, like study 1, was also semistructured and open ended. Note that study staff conducting the interviews were different between study 1 and study 2 (which also favors generalizability and has important ecological relevance).

2.3. Interview preprocessing and initial content analyses

All interviews were recorded in-person with an electronic handheld device and later transcribed to text. These interviews were preprocessed and analyzed using a pipeline detailed elsewhere.<sup>4</sup> For brevity, here we report only a summary of the relevant methodology.

In our previous work, we reported the analyses of 348 language features, using a cross-validated machine-learning pipeline that enforces sparsity with least absolute shrinkage and selection operator regularization.<sup>4</sup> In this study, we only examined the subset of language features that were identified and predictive in the previous article (Fig. 1C). These are 3 language features from Linguistic Inquiry Word Count (LIWC, version 2015<sup>33</sup>), namely the number of occurrences of words semantically associated with “Drives,” “Achievement,” and “Leisure,” and 8 features from semantic proximity metrics, namely semantic distance to “Magnify,” “Afraid,” “Fear,” “Awareness,” “Loss,” “Identity,” “Stigma,” and “Force.”

Linguistic Inquiry Word Count labels the words from a given text into semantic and syntactic categories, providing a measure of how frequently certain categories were used by the participant

during the interview (normalized for word count).<sup>33</sup> We also extracted semantic proximity metrics using Latent Semantic Analyses, see Ref. 24. Semantic proximity is a measurement of how close one word is to another in semantic space. This approach takes advantage of the fact that semantically related words tend to co-occur frequently. To do this, we extracted all words from the Touchstone Applied Science Associates (TASA) collection, which compiles thousands of text documents representing common knowledge across the U.S. educational system and generated a co-occurrence matrix for all pairs of words; this matrix was then reduced to 300 latent features using singular value decomposition. Each word can now be represented by a vector corresponding to the value for each of the 300 latent variables. This effectively maps each word from TASA into a semantic space. Semantic proximity is quantified by the dot product between the vectors between 2 given words. The more similar they are in semantic space, the higher the dot product, and a dot product of zero means that they are orthogonal or unrelated. We calculated the dot product between each word the patient used during the interview and the 8 topics of interest mentioned above. These dot products are then averaged providing a measure of how semantically proximal the whole interview is to these topics of interest. Because a shorter interview will inevitably result in patients using less words, it is more likely that the average semantic distances of the interview are affected by outlier words. Thus, semantic distances for the study 2 interview were calculated using the median value, instead of the mean. This was decided before any data analysis, and analyses with means



were never conducted. Furthermore, to ensure that the data were appropriately scaled for both interviews such that model parameters could be generalized, language features from study 2 were scaled according to study 1 data using a robust scaling method (normalization by interquartile range). More details regarding the scientific rationale and the methods applied here are presented in our previous article.<sup>4</sup>

## 2.4. Defining a placebo responder

For both studies and regardless of the treatment arm, participants were stratified into responders and nonresponders based on a permutation test of their pain ratings acquired during baseline against those acquired during the treatment periods (**Fig. 1A**): The null distribution was generated by shuffling 10,000 times pain ratings in the baseline and the treatment periods and then by comparing the new rating rearrangements at each iteration. T tests were used to determine whether the baseline and pain rating treatments differed significantly ( $P < 0.05$ ), in which the subject was labeled as a responder. Otherwise, subjects are classified as nonresponders. Please note that although this criterion could reflect other changes in patients' pain caused by regression to the mean and spontaneous recovery, here, we use a no-treatment arm as a control to specifically assess whether the model predicts placebo effects caused by the inert pill.

## 2.5. Model building and validation approach

In our previous work, we have demonstrated the ability of language quantitative features to identify placebo responders using a nested cross-validation approach.<sup>4</sup> This approach provides the opportunity to study accuracy within a small sample without overfitting the data, especially in a small  $n$  to large features scenario. The downside of this approach is that it does not provide us with 1 single model from which to predict from, but as many models as there are subjects (in the Leave-One-Out cross-validation case, which we used). To overcome this and generate a single model and like previous work,<sup>36</sup> we selected the words that were identified by the nested Leave-One-Out cross-validation model in the previous article and built a new, single model from study 1 data, using a logistic regression. Because there were 11 features and we observed evidence of collinearity among them (supplementary Figure 2, available at <http://links.lww.com/PAIN/B734>), we further used bidirectional stepwise selection (ie, a combination of forward and backward elimination after a  $P < 0.05$  threshold) to reduce our model into as few parameters as possible, to prevent overfitting, and improve generalization (**Fig. 1D**).

To validate the study 1 model, we used the linear equation from the study 1 model to generate predictions in study 2 data (**Fig. 1E**). Corresponding area under the curves (AUC) were generated with the results from the prediction model. Statistical significance was assessed by permuting responder labels 5000 times and using this null distribution to calculate  $P$  values; 95% confidence intervals for AUCs were obtained by bootstrapping with 5000 iterations. Balanced accuracy (to adjust for possible class imbalances), F1-score, precision, and recall are also reported, using a fixed cut off where subjects with a predicted probability  $> 0.5$  (range 0-1) are labeled as responders and  $\leq 0.5$  as nonresponders, as common in binary classification problems.

To explore the predictive ability of each feature independently, including those not in the logistic stepwise model, we performed univariate analyses. To do so, each of the 11 a priori selected features was fit to study 1 data and validated in study 2 data.

## 2.6. Content analyses

To further explore the language content associated with the latent language features *force*, *magnify*, and *stigma*, we traced the semantic distance properties back to the subject's original interview. To do so, each word in the interview was ranked by the semantic distance to the features *force*, *magnify*, *stigma*. Then, the top 5 words for each subject were extracted and counted for frequency across all subjects. Only words appearing at least twice were kept. These were used to construct word clouds which identify frequent words used in the interview. This also allows us to compare the specific words that the subjects used across study 1 and study 2 to (qualitatively) assess whether the subjects are using similar words and descriptors despite the differences in interviews. Text excerpts from the interview of patients that score the highest for each latent semantic category were collected for illustration purposes.

## 3. Results

### 3.1. Participants and outcomes

This article analyzes data from 2 longitudinal studies, examining the placebo response to an inert pill in patients with chronic low back. The first study (study 1, **Fig. 1A**) was designed to generate and tune placebo prediction models. Here, 66 patients completed all aspects of this study including the language interview. Of these patients, 4 received active treatment for blinding purposes and were excluded. The final sample for study 1 thus consists of 62 patients with CBP (mean =  $45.3 \pm 2.3$  years), 20 of which were assigned to receive no treatment. The remaining 42 subjects were assigned to the placebo group: 23 received significant pain relief (ie, significant change from baseline) from placebo and 19 did not (henceforth placebo responders and nonresponders, respectively; 55% responders). For study 2 (occurring over a year after study 1 completion in an entirely different set of patients), we witnessed a large attrition rate: only 46 patients completed all aspects of this study (**Fig. 1B**). Four patients were assigned to a no-treatment arm and were excluded from the analyses. Of the remaining 42 patients (mean age =  $45.3 \pm 2.3$  years), 20 were randomized to receive a placebo and 22 to receive naproxen + esomeprazole (active treatment, ie, drug group). For the placebo group, 8 were responders and 12 nonresponders (43% responders). For the drug group, 15 were responders and 7 were nonresponders (67% responders), see **Figure 1B**. Magnitudes of analgesia treatment arm by outcome are presented in supplementary Figure 3 (available at <http://links.lww.com/PAIN/B734>).

### 3.2. Model generation (initial data set, study 1)

To generate a single predictive model, we took a set of 11 features identified a priori based on previous work (Ref. 13, **Fig. 1C**) and fit them with a bidirectional stepwise logistic regression. The final model retained 4 of the initial 11 features (**Fig. 1D**): "achievement" from LIWC and semantic proximity to "force," "stigma," and "magnify." This model was highly accurate, being able to identify placebo responders at 91% accuracy and showing an AUC of 0.96,  $P < 0.001$ , reaching almost a perfect separation between placebo responders and nonresponders. Naturally, given that these features were selected on top of a set of already highly predictive features, the classification accuracy is optimistic (unbiased cross-validated accuracy is 79%, see Ref. 4). The logistic equation, with intercept and the 4 coefficients, as shown in **Figure 1D**, was used to predict placebo response in the second study.

### 3.3. Model validation (independent data set, study 2)

We applied the study 1 model to the data collected from study 2, and AUCs were calculated to assess predictive performance. Within the placebo group, the model predicted that 11 of 20 (55%) patients were placebo responders and within the drug group the model predicted that 17 of 23 (77%) patients were placebo responders.

For the placebo arm, this model showed good classification accuracy with an AUC = 0.708 (95% CI: 0.460–0.957),  $P = 0.054$  (Fig. 2A). As can be inspected in the confusion matrix (Fig. 2A, upper right panel), this model showed an F1-score = 0.65, precision score = 0.68, a recall score = 0.65, and a balanced accuracy of 67%. Subjects whom the model predicted as placebo responders showed higher magnitudes of analgesia compared with predicted nonresponders (30% vs 3% reduction in pain, respectively,  $P = 0.049$ ), with a large effect size (Hedge  $g = 0.90$ , see also Figure S4, <http://links.lww.com/PAIN/B734>). The same model applied to the drug treatment group (naproxen) showed unsatisfactory classification accuracy, with an AUC = 0.516 (95% CI: 0.283–0.760),  $P = 0.43$  (Fig. 2C). Classification metrics were F1-score = 0.62, precision score = 0.61, and recall score = 0.64, for a balanced accuracy of 54%. Although the mean analgesia of predicted responders (23%) was larger than predicted nonresponders (7%), this difference was not statistically significant ( $P = 0.19$ , Hedge  $g = 0.66$ , see also Figure S4, <http://links.lww.com/PAIN/B734>) probably because of the small number of predicted nonresponders,  $n = 5$ , see Figure 2C. Finally, if the 2 groups were analyzed together (patients were blinded to which pill they were taking; therefore, naproxen reasonably be expected to lead to placebo analgesia), the model showed satisfactory classification accuracy, with an AUC = 0.661 (95% CI: 0.512–0.796),  $P = 0.039$ . Classification metrics were F1-score = 0.63, precision score = 0.64, and recall score = 0.64, for a balanced accuracy of 63%.

To further account for the possibility that the model is predicting regression to the mean and spontaneous recovery, we further examined the ability of the model to predict pain relief in the no-treatment arm in both studies. Owing to the small number of subjects, we combined the subjects from both studies ( $N = 24$ ). The model provided unsatisfactory predictive ability (AUC = 0.55), which is consistent with the idea that it is indeed predicting placebo effects.

### 3.4. Univariate prediction (from study 1 to study 2)

For completeness, we also examined the predictive ability of the features not included in the main model. To do so and independently for each feature, we fitted a logistic regression on study 1 data and tested it in study 2 data. For the placebo group, most features show acceptable classification accuracy (AUC > 0.6), with semantic proximity to awareness showing the highest single predictive ability (AUC = 0.69). By contrast, for the main treatment group, most features showed poor classification accuracy (AUCs < 0.6, except for identity, AUC = 0.62). Interestingly, semantic proximity to fear was able to correctly misclassify drug responders with an AUC of 0.27.

### 3.5. Extracting meaning: word associations of the original interviews

To further probe the meaning behind the latent semantic topics, we traced back the semantic distance features to the patient's original discourse. Word frequency clouds for each semantic distance feature can be inspected in Figure 3 as well as some illustrative sentence-level examples. Words associated with

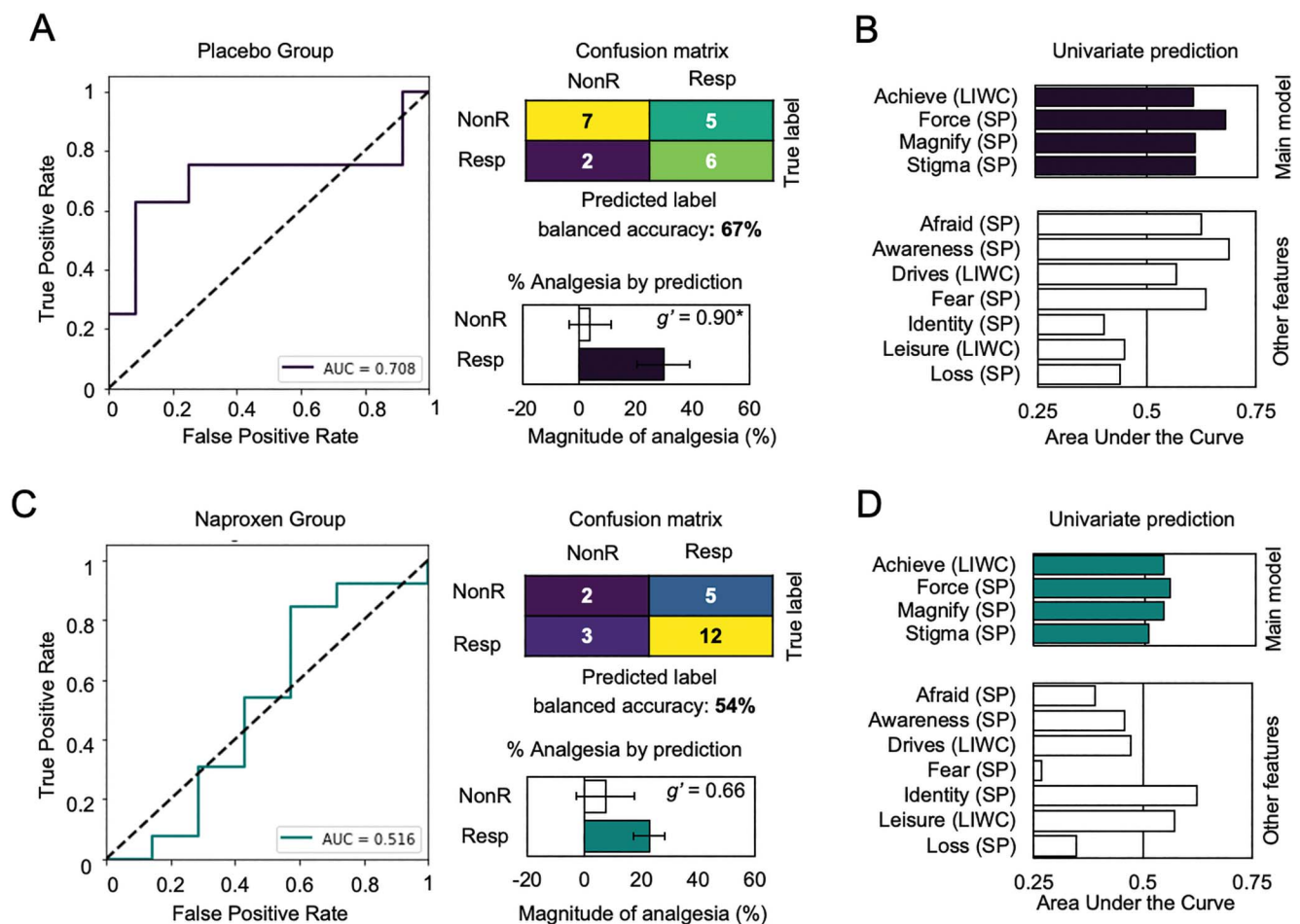
“force” are associated with physical forces such as pull, push, lift, and rest and appear equally in both study 1 and study 2 interviews. Other features are harder to interpret in isolation: Common words for *magnify* include “describe,” “kind,” “real,” “sharp,” and “x-ray” and *stigma* was associated with words such as “long,” “another,” “call,” and “ever.”

## 4. Discussion

We report the results from a validation study of a placebo prediction model based on features derived from natural language processing. Consistent with our previous findings,<sup>4</sup> a logistic model built on latent semantic features from an open-ended interview in a first study was able to successfully identify placebo responders from nonresponders with good accuracy in a new independent study (AUC = 0.71). Although the sample size was small and the statistical power was low, considering the model was trained and validated on data sets with different interview lengths (first study: 27 minutes and 16 questions and second study: 3 minutes and 4 questions), at a different time point (first study: end of the treatment and second study: before treatment commencement), and elicited by different interviewers, we take these findings as compelling evidence of the robustness and generalizability of the model.

Within individuals, it is likely that the placebo response will be determined not only by stable characteristics of the individual such as personality and brain properties<sup>36–38</sup> but also by context, expectations, previous experiences, and even the type of placebo administered.<sup>2,10,18,27,42</sup> Thus, arguably, a good way to approach the prediction of placebo analgesia is by using methods that can explore these traits in an ecological manner. This implies understanding where the subjects' pain comes from, how they deal with it, their expectations, their frustrations with the previous medical experiences, and, more generally, how these patients see themselves, others, and the world.<sup>29</sup> Examining patients' language content through an open-ended interview is thus a powerful tool because it provides information linked with the person's subjective and unique experiences.<sup>4,22</sup> Here, we show that patterns in language use can be quantified and used to identify patients who may benefit from pain analgesia from an inert pill. In fact, patients who were predicted as placebo responders before the treatment had an average magnitude of analgesia of 30%, an amount that is clinically significant, vs an average of 3% pain analgesia for nonresponders. Importantly, we have shown that brain and personality measures can be effective at predicting placebo, but with suboptimal ability to classify individual subjects.<sup>36</sup> The fact that this language model seems to perform better than a model derived from brain and personality features indeed suggests that the placebo response, at least at the individual level, is best predicted from a more ecological approach. This could hinge on the fact that personality measures tend to focus on more stable traits, which downplays both current state of the patient as well as its immediate context and past experiences. We speculate that assessing patients' experiences quantitatively through language may tap into psychological and psychosocial dimensions that are not easily accessible by conventional psychometric approaches or may be obfuscated by them.

The model was able to predict placebo responders, but not drug responders, showing specificity of prediction. This result is surprising because we expected some amount of pain analgesia in the drug group to be caused by placebo additive effects.<sup>36</sup> This is, however, convincing evidence that the model is not predicting some trivial property caused by regression to



**Figure 2.** Placebo predictive model validates in an independent sample. (A) Study 1 logistic model validates in study 2 data for the placebo treatment group, predicting placebo responders with an AUC of 0.71. Right upper panel shows the confusion matrices from the original model predictions, which resulted in a balanced accuracy of 67%. The actual magnitude of analgesia observed in predicted placebo responders was 30% and significantly larger than those predicted as nonresponders (3%). (B) Univariate, all the features in the main model showed above chance predictability (AUCs > 0.6); other features selected a priori show equally good predictability, with semantic proximity to awareness showing the highest overall AUC (0.69). (C) By contrast, the placebo model does not predict response to drug treatment with a poor predictive performance (AUC = 0.52). Confusion matrices show that the model was able to predict responders quite effectively but, given how the sample is heavily unbalanced, this led to a poor balanced accuracy of 54%. The actual magnitude of pain analgesia was larger but not statistically significant between predicted responders and nonresponders (27 vs 7%,  $P = 0.17$ ). (D) Unlike in the placebo group, univariate features show poor AUCs, except for semantic proximity for identity predicting drug responders with an AUC of 0.65. AUC, area under the curve; LIWC, Linguistic Inquiry Word Count; NonR, non-responder; Resp, responder; SP, semantic proximity.

the mean or natural history effects because these should be equivalent across treatment types. Despite poor predictability at the single-subject level, predicted responders did have tendency of more analgesia than those predicted as nonresponders, and in fact, the model successfully predicted drug responders quite accurately (12 of 15 or 80%) but failed to identify nonresponders (2 of 7 or 29%). We hypothesize that predicted placebo responders who responded to the drug would have had larger magnitudes of analgesia than those who were predicted as nonresponders but responded to the drug, but unfortunately, the large percentage of drug responders, the small sample size, and the even smaller number of predicted nonresponders preclude us from drawing conclusions about this effect. Of course, because the model was specifically trained in a data set that only included placebo-treated patients, it is not particularly tailored to predict drug responses. A better model could have been built which was specifically designed to predict drug responses or treatment outcomes, an interesting concept that motivates further studies.

In both study 1 and study 2 and consistent with our previous report,<sup>4</sup> patients whose answers are semantically closer to “force”

and “stigma” are less likely to respond to placebo. An examination of the words patients used in both studies shows that “force” is related to how patients describe physical forces and their relationship with pain (eg, “pull,” “push,” “lift,” and “effort”). It is tempting to suggest that patients who use these words to describe their pain are less likely to respond to placebo pills because the source of their pain is clearly defined and expected, a hypothesis that is supported by current predictive models of placebo analgesia.<sup>5</sup> Words associated with “stigma” do not allow a straightforward interpretation (most frequent words are “long,” “call,” and “another”), yet a qualitative examination of the subjects discourse reflect patients who, for multiple and heterogeneous reasons, lack trust in the medical system or feel they are stigmatized (eg, “it is not fair to just say the reason that you are like that is because you are fat”). It has been shown that previous therapeutic experiences predict placebo effects,<sup>9</sup> so patients feeling stigmatized might have lesser expectations of getting pain relief from medical care (and a placebo, for that matter<sup>2</sup>), which is in line with research on placebo expectations reducing placebo effectiveness.<sup>10</sup>

In the opposite direction, patients with a higher number of words tagged under “achievement” and higher semantic





**Figure 3.** Content analysis and text excerpts. Words highly associated with the 3 semantic distance features were captured and quantified by frequency for study 1 and study 2. Word clouds show words that appeared at least twice throughout the interviews, size-scaled by frequency. Red colors denote words that appeared in both study 1 and study 2. For the 3 features, there are common words that are being used, showing that the semantic distance metrics are mapping similar topics. Illustrative examples for each feature and study are given (middle row).

proximity to “magnify” are more likely to respond to placebo. In both studies, “magnify” was linked with words such as “real,” “x-ray,” “sharp,” and “describe” and may reflect an increased focused of attention (ie, magnifying), pain, and bodily sensations (eg, “my pain is hard to quantify but when it hits you, oh my god”). Previous studies have linked interoceptive awareness as a predictor of placebo analgesia<sup>38</sup> and somatic focus as a promoter of placebo effects<sup>16,17</sup>; this also fits in well with our post hoc finding that semantic proximity to “awareness” can itself identify placebo responders with good accuracy (AUC = 0.69). Similarly, “achievement” words counted with LIWC could be related to motivation as well as the ability to do work, be in control,

and act proactively to obtain pain relief and seek care. Naturally, the data-driven approach used here lends itself to speculative explanations, so proper and justified interpretation of these language features requires future studies.

Another point to discuss is that this study was conducted in a patient population of patient with CBP, where the motivation and expectation to get pain relief from their chronic condition could be higher than in the laboratory setting. In fact, pain conditions show some of the largest placebo effects in clinical trials,<sup>20</sup> and it has been shown that larger pain intensity leads to higher placebo efficacy.<sup>25</sup> We argue that this favors the predictability of the effect because expectations are believed

to be important for placebo,<sup>10,26,29</sup> but see Ref. 21. Although in our studies, patients were unbiased to expectations, that is, they were told that they might receive a placebo or a drug with no indication of active treatment likelihood, the placebo analgesia found in both our studies is quite substantial and long lasting—the whole group receiving a placebo pill experienced 21% and 22% average pain reduction across study 1 and 2, respectively, and the subgroup of placebo responders showed average pain reductions of 32% and 30%, respectively. The large and clinically significant pain reduction found here supports the view that the placebo is more pronounced, and perhaps more predictable, in the clinical context<sup>9,15</sup>—and particularly for chronic low back pain. In fact, given the success found in predicting placebo responses in CBP, it is now necessary to understand how applicable these results are to other chronic pain conditions. Furthermore, given that participants were provided with neutral instructions, this analgesia magnitude could be even further increased through the manipulation of analgesic expectations<sup>12,30,32</sup> or by using more invasive placebo treatments than pills.<sup>27,42,43</sup>

Finally, and more generally, this study further demonstrates the power of language to study behavior broadly,<sup>33</sup> and given how reliable and easy a short interview is to implement, it opens new avenues to study and predict treatment and drug responses in other clinical conditions, a field that only recently has received attention.<sup>1,6,11,28</sup> In addition, the work here was constrained by backward compatibility with our previous work.<sup>4</sup> Recent advances in natural language processing, including transformed-based models such as BERT<sup>13,31</sup> which account for the semantic nuances implied by the context of single target words in sentences and paragraphs, are superior to bag-of-words methods as used in this study, at the cost of requiring significantly more data; these might further improve the accuracy of placebo prediction and provide more contextually relevant and easier to interpret language features.

This study has some limitations. First, the sample size is small, which led to a marginally significant classification accuracy for the placebo group ( $P = 0.054$ ) as well as wide 95% confidence intervals (0.45–0.97); this precludes us from making strong claims regarding the true predictability of the model. Second, the interview lengths are not matched between study 1 and study 2, and the interviews were collected at different time points. We speculate that if the methods were comparable, the classification accuracy could have been improved. Finally, because of limited data in both studies, we used relatively simple natural language processing models; new studies should explore more state-of-art approaches such as Bidirectional Encoder Representations from Transformers (ie, BERT).

In summary, our results support the thesis that placebo response is predictable and can be examined objectively through the study of mental processes that are, as shown here, reflected onto the semantic content of patients' speech. That language features dissociate placebo responders from nonresponders have important implications not only for clinical practice but also for study designs. For instance, identifying placebo responders has the potential of improving clinical trial design, allowing for a more efficient allocation of participants to treatment arms (with equal predicted responders in all arms), discounting the placebo effect size parametrically, or eliminating the placebo confound altogether by excluding predicted responders during enrollment. Larger-scale studies are now necessary to further assess generalizability and precisely estimate the true accuracy of this predictive model.

## Conflict of interest statement

The authors have no conflicts of interest to declare.

## Acknowledgments

The authors would like to thank all members of the Apkarian Lab for their feedback on the manuscript and three anonymous reviewers for their constructive feedback. This work was funded by the National Center for Complementary and Integrative Health AT007987 and National Institutes of Health Grant P50 DA044121. E. Vachon-Preseu was funded through Canadian Institutes of Health Research (CIHR). The authors declare no competing interests.

## Appendix A. Supplemental digital content

Supplemental digital content associated with this article can be found online at <http://links.lww.com/PAIN/B734>.

## Article history:

Received 25 February 2022

Received in revised form 29 September 2022

Accepted 5 October 2022

Available online 18 October 2022

## References

- [1] Agurto C, Cecchi GA, Norel R, Ostrand R, Kirkpatrick M, Baggott MJ, Wardle MC, de Wit H, Bedi G. Detection of acute 3, 4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing. *Neuropsychopharmacology* 2020;45:823–32.
- [2] Atlas LY. A social affective neuroscience lens on placebo analgesia. *Trends Cogn Sci* 2021;25:992–1005.
- [3] Benedetti F. The opposite effects of the opiate antagonist naloxone and the cholecystikinin antagonist proglumide on placebo analgesia. *PAIN* 1996;64:535–43.
- [4] Berger SE, Branco P, Vachon-Preseu E, Abdullah TB, Cecchi G, Apkarian AV. Quantitative language features identify placebo responders in chronic back pain. *PAIN* 2021;162:1692–704.
- [5] Büchel C, Geuter S, Sprenger C, Eippert F. Placebo analgesia: a predictive coding perspective. *Neuron* 2014;81:1223–39.
- [6] Carrillo F, Sigman M, Fernández Slezak D, Ashton P, Fitzgerald L, Stroud J, Nutt DJ, Carhart-Harris RL. Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. *J Affective Disord* 2018;230:84–6.
- [7] Carvalho C, Caetano JM, Cunha L, Rebouta P, Kaptchuk TJ, Kirsch I. Open-label placebo treatment in chronic low back pain: a randomized controlled trial. *PAIN* 2016;157:2766–72.
- [8] Carvalho C, Pais M, Cunha L, Rebouta P, Kaptchuk TJ, Kirsch I. Open-label placebo for chronic low back pain: a 5-year follow-up. *PAIN* 2021;162:1521–7.
- [9] Colloca L, Akintola T, Haycock NR, Blasini M, Thomas S, Phillips J, Corsi N, Schenk LA, Wang Y. Prior therapeutic experiences, not expectation ratings, predict placebo effects: an experimental study in chronic pain and healthy participants. *Psychother Psychosomat* 2020;89:371–8.
- [10] Corsi N, Colloca L. Placebo and nocebo effects: the advantage of measuring expectations and psychological factors. *Front Psychol* 2017;8:308.
- [11] Cox DJ, Garcia-Romeu A, Johnson MW. Predicting changes in substance use following psychedelic experiences: natural language processing of psychedelic session narratives. *Am J Drug Alcohol Abuse* 2021;47:444–54.
- [12] De Pascalis V, Chiaradia C, Carotenuto E. The contribution of suggestibility and expectation to placebo analgesia phenomenon in an experimental setting. *PAIN* 2002;96:393–402.
- [13] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* 2019:1810.04805 [cs]. Available at: <http://arxiv.org/abs/1810.04805>. Accessed November 17, 2021.



- [14] Finnerup NB, Haroutounian S, Baron R, Dworkin RH, Gilron I, Haanpaa M, Jensen TS, Kamerman PR, McNicol E, Moore A, Raja SN, Andersen NT, Sena ES, Smith BH, Rice AS, Attal N. Neuropathic pain clinical trials: factors associated with decreases in estimated drug efficacy. *PAIN* 2018; 159:2339–46.
- [15] Forsberg JT, Martinussen M, Flaten MA. The placebo analgesic effect in healthy individuals and patients: a meta-analysis. *Psychosom Med* 2017; 79:388–94.
- [16] Geers AL, Helfer SG, Weiland PE, Kosbab K. Expectations and placebo response: a laboratory investigation into the role of somatic focus. *J Behav Med* 2006;29:171–8.
- [17] Geers AL, Wellman JA, Fowler SL, Rasinski HM, Helfer SG. Placebo expectations and the detection of somatic information. *J Behav Med* 2011;34:208–17.
- [18] Horing B, Weimer K, Muth ER, Enck P. Prediction of placebo responses: a systematic review of the literature. *Front Psychol* 2014;5:1079.
- [19] Hróbjartsson A, Gotzsche PC. Is the placebo powerless? *N Engl J Med* 2001;344:1594–602.
- [20] Hróbjartsson A, Gotzsche PC. Placebo interventions for all clinical conditions. *Cochrane database Syst Rev* 2010:CD003974. doi: 10.1002/14651858.CD003974.pub3.
- [21] Kaptchuk TJ, Hemond CC, Miller FG. Placebos in chronic pain: evidence, theory, ethics, and use in clinical practice. *BMJ* 2020;370:m1668.
- [22] Kaptchuk TJ, Shaw J, Kerr CE, Conboy LA, Kelley JM, Csordas TJ, Lembo AJ, Jacobson EE. “Maybe I made up the whole thing”: placebos and patients’ experiences in a randomized controlled trial. *Cult Med Psychiatry* 2009;33:382–411.
- [23] Kleine-Borgmann J, Schmidt K, Hellmann A, Bingel U. Effects of open-label placebo on pain, functional disability, and spine mobility in patients with chronic back pain: a randomized controlled trial. *PAIN* 2019;160:2891–7.
- [24] Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. *Discourse Process* 1998;25:259–84.
- [25] Levine JD, Gordon NC, Bornstein JC, Fields HL. Role of pain in placebo analgesia. *Proc Natl Acad Sci U S A* 1979;76:3528–31.
- [26] Linde K, Witt CM, Streng A, Weidenhammer W, Wagenpfeil S, Brinkhaus B, Willich SN, Melchart D. The impact of patient expectations on outcomes in four randomized controlled trials of acupuncture in patients with chronic pain. *PAIN* 2007;128:264–71.
- [27] Meissner K, Fässler M, Rücker G, Kleijnen J, Hróbjartsson A, Schneider A, Antes G, Linde K. Differential effectiveness of placebo treatments: a systematic review of migraine prophylaxis. *JAMA Intern Med* 2013;173:1941–51.
- [28] Norel R, Agurto C, Heisig S, Rice JJ, Zhang H, Ostrand R, Wacnik PW, Ho BK, Ramos VL, Cecchi GA. Speech-based characterization of dopamine replacement therapy in people with Parkinson’s disease. *Npj Parkinsons Dis* 2020;6:12–8.
- [29] Price DD, Finniss DG, Benedetti F. A comprehensive review of the placebo effect: recent advances and current thought. *Annu Rev Psychol* 2008;59:565–90.
- [30] Reicherts P, Pauli P, Möslers C, Wieser MJ. Placebo manipulations reverse pain potentiation by unpleasant affective stimuli. *Front Psychiatry* 2019; 10:663.
- [31] Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works. *Trans Assoc Comput Linguist* 2020;8: 842–66.
- [32] Stewart-Williams S, Podd J. The placebo effect: dissolving the expectancy versus conditioning debate. *Psychol Bull* 2004;130:324–40.
- [33] Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol* 2010;29:24–54.
- [34] Tétrault P, Mansour A, Vachon-Presseau E, Schnitzer TJ, Apkarian AV, Baliki MN. Brain connectivity predicts placebo response across chronic pain clinical trials. *PLoS Biol* 2016;14:e1002570.
- [35] Tuttle AH, Tohyama S, Ramsay T, Kimmelman J, Schweinhardt P, Bennett GJ, Mogil JS. Increasing placebo responses over time in U.S. clinical trials of neuropathic pain. *PAIN* 2015;156:2616–26.
- [36] Vachon-Presseau E, Abdullah TB, Berger SE, Huang L, Griffith JW, Schnitzer TJ, Apkarian AV. Validating a biosignature predicting placebo pill response in chronic pain in the settings of a randomized controlled trial. *PAIN* 2021;163:910–22.
- [37] 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.
- [38] Vachon-Presseau E, Berger SE, Abdullah TB, Huang L, Cecchi GA, Griffith JW, Schnitzer TJ, Apkarian AV. Brain and psychological determinants of placebo pill response in chronic pain patients. *Nat Commun* 2018;9:3397.
- [39] Vase L, Riley JL, Price DD. A comparison of placebo effects in clinical analgesic trials versus studies of placebo analgesia. *PAIN* 2002;99: 443–52.
- [40] Vase L, Robinson ME, Verne NG, Price DD. Increased placebo analgesia over time in irritable bowel syndrome (IBS) patients is associated with desire and expectation but not endogenous opioid mechanisms. *PAIN* 2005;115:338–47.
- [41] 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.
- [42] Zhang W, Robertson J, Jones AC, Dieppe PA, Doherty M. The placebo effect and its determinants in osteoarthritis: meta-analysis of randomised controlled trials. *Ann Rheum Dis* 2008;67:1716–23.
- [43] Zou K, Wong J, Abdullah N, Chen X, Smith T, Doherty M, Zhang W. Examination of overall treatment effect and the proportion attributable to contextual effect in osteoarthritis: meta-analysis of randomised controlled trials. *Ann Rheum Dis* 2016;75:1964–70.
- [44] Zunhammer M, Spisák T, Wager TD, Bingel U. Meta-analysis of neural systems underlying placebo analgesia from individual participant fMRI data. *Nat Commun* 2021;12:1391.