Research Paper

PAIN®

VIDEO

Quantitative language features identify placebo responders in chronic back pain

Sara E. Berger[a,b], Paulo Branco[a], Etienne Vachon-Presseau[c,d,e], Taha B. Abdullah[a], Guillermo Cecchi[b], A. Vania Apkarian[a,f,g,h,*]

Abstract

Although placebo effect sizes in clinical trials of chronic pain treatments have been increasing, it remains unknown if characteristics of individuals' thoughts or previous experiences can reliably infer placebo pill responses. Research using language to investigate emotional and cognitive processes has recently gained momentum. Here, we quantified placebo responses in chronic back pain using more than 300 semantic and psycholinguistic features derived from patients' language. This speech content was collected in an exit interview as part of a clinical trial investigating placebo analgesia (62 patients, 42 treated; 20 not treated). Using a nested leave-one-out cross-validated approach, we distinguished placebo responders from nonresponders with 79% accuracy using language features alone; a subset of these features—semantic distances to identity and stigma and the number of achievement-related words—also explained 46% of the variance in placebo analgesia. Importantly, these language features were not due to generic treatment effects and were associated with patients' specific baseline psychological traits previously shown to be predictive of placebo including awareness and personality characteristics, explaining an additional 31% of the variance in placebo analgesia beyond that of personality. Initial interpretation of the features suggests that placebo responders differed in how they talked about negative emotions and the extent that they expressed awareness to various aspects of their experiences; differences were also seen in time spent talking about leisure activities. These results indicate that patients' language is sufficient to identify a placebo response and implie that specific speech features may be predictive of responders' previous treatment.

Keywords: Placebo response, Chronic back pain, Natural language processing, Interview, Semantic proximity, Machine learning, Psycholinguistics

1. Introduction

Chronic pain is a highly prevalent and poorly managed pathology that remains a mystery to the medical system and exacts a major socioeconomic burden.[9] There is no single treatment superior to others for relieving chronic pain, with available care providing only modest short-term improvements, and more aggressive options potentially leading to adverse events without sustained relief. Blinded randomized control trials (RCTs) consistently show that the placebo response is not only universally observed but also leads to substantial analgesia in chronic pain trials.[43] Recent evidence indicates that placebo responders and nonresponders with chronic back pain (CBP) can be predicted from neurobiological and psychological parameters, for placebo pills[45] and for placebo patches.[18] Moreover, brain functional network properties can dissociate active treatment responses from placebo treatment responses in patients with chronic pain.[7,12,40,46] Nevertheless, much of the placebo response in the context of chronic pain remains minimally understood, including how to best measure it and what features of a patient's experience, personality, or thought processes comprise it.

Research in the fields of computational psychiatry and linguistics suggests that patients' language—ie, *what* they say and *how* they say it—can provide important psychosocial signals.[16,39] An interest in using language to investigate human behavior has gained momentum because of its neurobiological roots and intimate role in the construction of thoughts, emotions, and experience—language is a window into the mind.[32] In recent years, researchers have used language features to measure personality and infer mood,[20,31–33,36,42] predict onset of psychological disorders,[1,8,37] and identify classes and doses of ingested drugs.[2] These findings provide support for using language to not only better understand complex conditions but also to dissociate subgroups of patients and identify changes in health status because of treatment.

Previous studies have established that the placebo response in healthy subjects is driven by several psychological constructs including expectation bias, goal seeking, and self-esteem,[19] as

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

S.E. Berger and P. Branco contributed equally to this manuscript.

[a] Department of Physiology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [b] Health Care and Life Sciences Department, IBM Watson Research Center, Yorktown Heights, NY, United States, [c] Faculty of Dentistry, McGill University, Montreal, QC, Canada, [d] Department of Anesthesia, Faculty of Medicine, McGill University, Montreal, QC, Canada, [e] Alan Edwards Centre for Research on Pain (AECRP), McGill University, Montreal, QC, Canada, [f] Department of Physical Medicine and Rehabilitation, [g] Department of Anesthesia, and, [h] Center for Translational Pain Research, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

**Corresponding author. Address: Feinberg School of Medicine, Northwestern University, Tarry Bldg. 7-705, Chicago, IL 60611, United States. Tel.: (312) 503-0404; fax (312)503-5101. E-mail: 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).

PAIN 162 (2021) 1692–1704

© 2021 International Association for the Study of Pain

http://dx.doi.org/10.1097/j.pain.0000000000002175

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

well as by specific personality traits such as optimism, extraversion, and neuroticism.[15,30] Our earlier studies have also demonstrated that psychological factors including emotional awareness and openness are predictive of placebo response in patients with chronic pain.[44,45] Because quantitative language analyses provide a tool to directly tap into these mental constructs,[49] capturing patients' thoughts and emotions through a conversation not only makes scientific sense but also creates an opportunity to objectively quantify their experience in a more naturalistic way.

Here, we deploy quantitative methodologies to analyze language from patient interviews in the context of an RCT investigating clinical placebo analgesia in CBP. We hypothesized that placebo responders would show unique language patterns related to emotional processing, coping mechanisms, and previous experiences, and that these linguistic features would accurately dissociate responders from nonresponders. Considering the novelty and translational potential of this methodology to study the placebo response in patients, this study represents a *proof-of-concept* that language may be used to quantitatively measure and interpret complex aspects of the chronic pain experience and response to treatment.

2. Methods

2.1. Participants

The data were collected as part of a larger clinical trial investigating placebo in patients with chronic back pain (ClinicalTrials.gov registration ID: NCT02013427). Information about the study population and design is summarized here; more details about the trial (eg, visit activities, randomization block designs, or blinding procedures) can be found in the original article.[45]

One hundred twenty-nine participants with chronic back pain (CBP) were assessed for eligibility through an intake form completed with a study coordinator. All included participants were at least 18 years old, had a history of CBP for at least 6 months, had no evidence of comorbid pain, neurological, or psychological disorders, and had met all additional inclusion criteria found in our previous report. All participants stopped concomitant pain medications and reported an initial pain of at least 5 of 10 on a VAS scale before study commencement. The Northwestern University Institutional Review Board approved the study, and all participants signed an informed consent form at the beginning of visit 1.

2.2. Study design

The RCT consisted of 6 visits over 8 weeks; the study design is shown in **Figure 1A**. At visit 1, participants were screened for eligibility and provided with rescue medication (acetaminophen, 500 mg) to use if needed in place of their normal pain medication, which they discontinued for the study duration. They were trained to track their pain and mood on VAS scales twice a day for the study duration using a smartphone app. After a 2-week baseline rating period, participants returned at visit 2 and were randomized into either a no-treatment or treatment arm. Those assigned to the no-treatment allocation (N = 20) were used as a control group to account for natural fluctuations of pain and regression to the mean and followed for the duration of the study, completing identical study procedures as those in the treatment group with the exception of receiving a study medication. Those in the treatment group either received active treatment (naproxen, 500 mg + esomeprazole, 20 mg, b.i.d.) or placebo treatment (2 lactose pills, b.i.d.). All pills were identically encapsulated to ensure the maintenance of a double-blind between patients and research staff. Importantly, because the original trial was designed to investigate biomarkers and mechanisms of placebo response, most participants in the treatment group received placebo treatment (N = 42). The active treatment group (N = 4) was only used as a double-blinding tool, and their data were not analyzed here.

Patients completed two 2-week long treatment periods, each followed by a 1-week long washout when they discontinued their study medication. At the end of the study (visit 6), participants completed an exit interview designed to capture potential differences in language use between responders and nonresponders. A battery of questionnaires was completed at each visit (see supplementary material, available at http://links.lww.com/PAIN/B248). Participants were compensated for their time, travel expenses, and total phone ratings completed.

2.3. Defining a placebo response using app data

As previously described,[45] participants who received placebo treatment were stratified into placebo responders and nonresponders based on a permutation test of their pain ratings acquired during baseline against those acquired during either of the 2 treatment periods. The null hypothesis was generated by randomly shuffling the ratings in the baseline with the ratings in the treatment periods 10,000 times, statistically comparing the new rating rearrangements at each iteration. The *t*-value from the one-sample *t* test between baseline and treatment values was used to determine if the null hypothesis could be rejected for each treatment period. If the *t*-value fell within this distribution ($P \geq 0.05$) for both treatment periods, the patient was classified as a "nonresponder." Otherwise, the patient was classified as a "responder" (**Fig. 1B**). In a post hoc analysis, we investigated response as a continuous variable by subtracting the averaged baseline pain ratings from those entered during the last week of each treatment period. The "magnitude" of analgesia was the maximum percent difference between baseline and either of the 2 treatment periods. More details about data preprocessing and response definition rationale can be found in the Supplement (available at http://links.lww.com/PAIN/B248).

2.4. Blinding the analysis

After acquiring all the data, a laboratory member not involved in the analyses renamed files to blind researchers to the participant's unique ID; interview preprocessing was performed after IDs were blinded. To further reduce bias, 2 random spreadsheets with data were also generated. This was performed by randomly shuffling the placebo responder vs nonresponder labels. The 2 random lists and the real list were provided to the author performing the statistical analyses (P.B.), who remained blinded until the end of all analyses. Once the analyses were completed for the 3 lists, the results were presented publicly at a laboratory meeting and the blinding codes revealed.

2.5. Exit interview design and implementation

The interview script was piloted on 10 participants with knee osteoarthritis who had previously completed another clinical trial[40] to verify the content was appropriate and could elicit enough free speech. Briefly, the interview included a warm-up section (not pain or study related) and a main section that probed participants about their pain, emotions, medical experiences, and time in the study. Figure S1 shows the final interview script

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

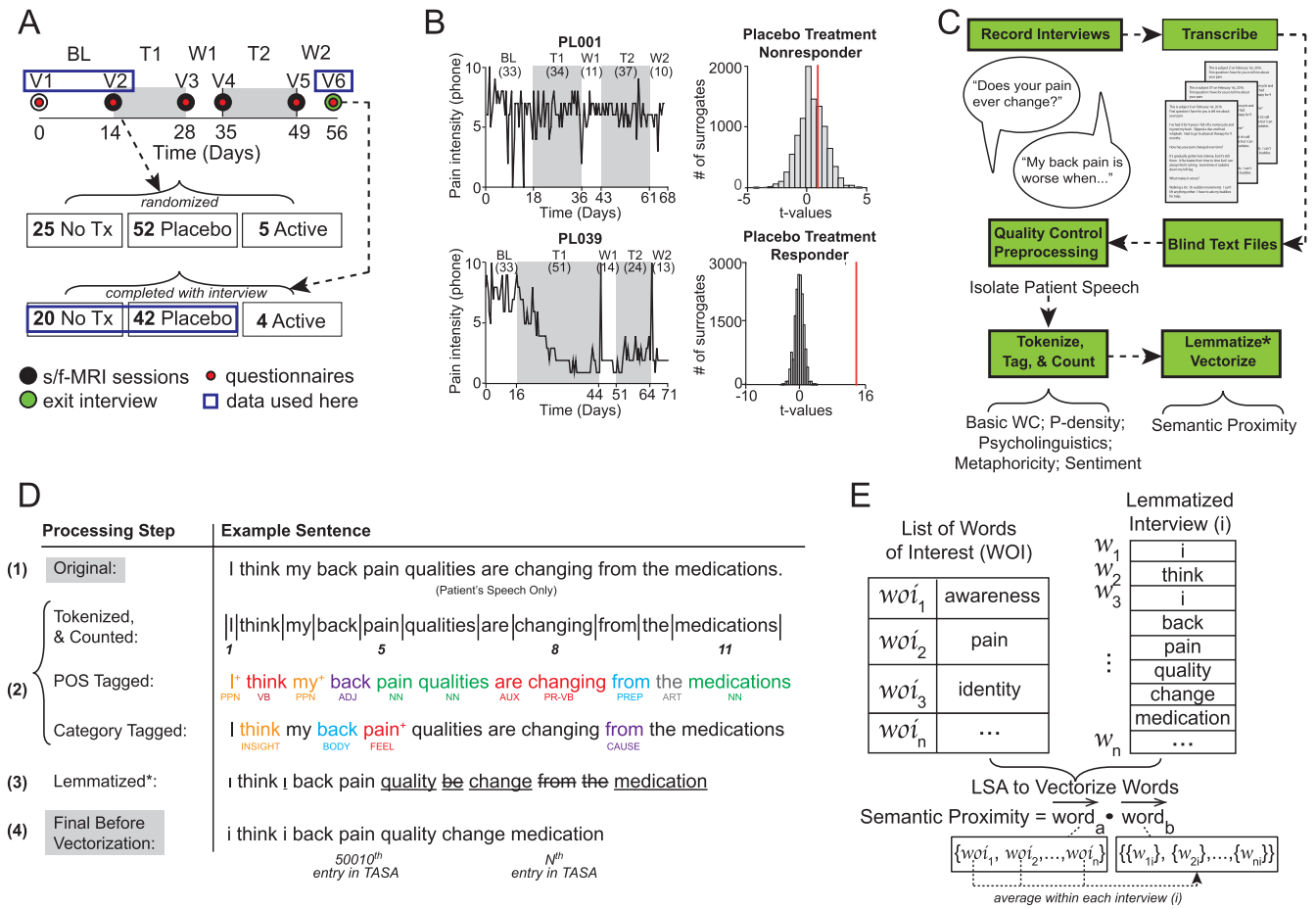


Figure 1. Study design and methodology. (A) Patients with chronic back pain (CBP) participated in 6 RCT visits over 8 weeks, completing a total of 4 brain scans (black dots), 2 consecutive treatment (T) and washout (W) periods, 6 questionnaire batteries (red dots), twice daily pain/mood ratings using a smartphone app, and a final exit interview (green dot). Outside of app ratings, only the data from the first, second, and sixth visit (blue boxes) are used in this article. (B) Time series show 2 patients' daily pain intensity ratings from baseline (BL) to end of the study. Gray panels indicate placebo treatment periods (T1 and T2), each followed by ~7-day washout (W1 and W2); # of ratings shown in parentheses. We binarized patients into responders and nonresponders using a permutation test. Red line indicates the real *t*-value between baseline and T1 or T2: if within distribution, a patient was labeled as a nonresponder; if outside distribution ($P < 0.05$), a patient was a responder. (A and B) Adapted from previous article[45] (C). After recording and transcription, interviews were blinded and patient speech extracted. The text was stripped and tokenized for specific analyses: word counts (WC), idea density (p-density), psycholinguistic inquiry (LIWC), metaphoricity, and sentiment. Stop words were removed; text was lemmatized and converted to numerical indices based on word location in the TASA corpus. Lemmas were vectorized using latent semantic analysis (LSA) to calculate semantic proximity to words of interest (WOIs). (D) An example of how a sentence changes during preprocessing. Phrases were tokenized (lines), and words were tagged across multiple dimensions; 2 examples—parts of speech and LIWC category—provided (+indicates application to multiple categories). Next, stop words were removed (strikethrough), and the remaining text was lemmatized. The final sentence before vectorization is shown (numbers = corpus location). (E) For each word in TASA, a corresponding 300-dimension vector was produced through latent semantic analysis (LSA) representing its semantic meaning within the English lexicon. The dot product between each interview word's vector and each of the 261 WOI vectors was calculated as a measure of semantic proximity. LIWC, linguistic inquiry and word count; RCT, randomized control trial; TASA, touchstone applied science associates; WOI, words of interest.

(supplemental methods explain the rationale for the interview setup, available at http://links.lww.com/PAIN/B248).

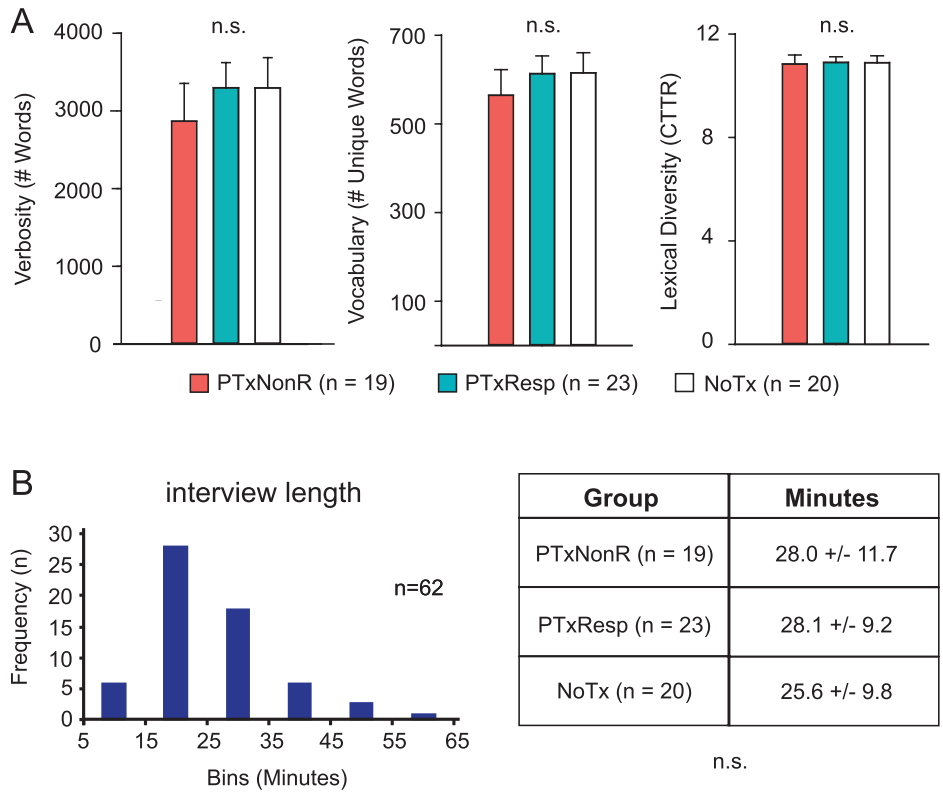
There is substantial evidence to suggest that future treatment expectations and perceived pain intensity can be influenced not only by emotions preceding an event[13] but also by how a researcher or practitioner talks to the patient[46]; moreover, it has been demonstrated that previous conversations can impact subsequent placebo and active treatment effect sizes.[14] A more detailed discussion of these effects can be found in the supplement (available at http://links.lww.com/PAIN/B248). Given this, all interviews were completed at the beginning of the last visit (visit 6), chosen to minimize any positive or negative emotions that may have influenced future placebo or nocebo responses had they occurred at visit 1.

For consistency between participants, only one researcher (S.E.B.) led interview questions for all participants; however, to control for potential sex effects,[29,34] a male colleague with whom the participant was familiar was also present in the room as an observer (E.V.P. or T.B.A.).

2.6. Interview preprocessing and initial content analyses

All interviews were recorded with an electronic handheld device and transcribed by an external company; transcription and quality control requirements can be found in the supplement (available at http://links.lww.com/PAIN/B248). After transcription, interviews were preprocessed in a multistep procedure graphically summarized in **Figures 1C and D**. A total of 6 analyses were conducted over 4 steps to calculate 348 language features: basic word counts (3 features), idea density (1 feature), psycholinguistic inquiry (78 features), metaphoricity (1 feature), sentiment (4 features), and semantic proximity (261 features) were extracted from the final preprocessed interview.

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.



Group	Minutes
PTxNonR (n = 19)	28.0 +/- 11.7
PTxResp (n = 23)	28.1 +/- 9.2
NoTx (n = 20)	25.6 +/- 9.8

Figure 2. No differences in lower-level language features between no-treatment, placebo responders, and nonresponders. (A) Verbosity, vocabulary, and lexical diversity were not different between PTxResp, PTxNonR, and NoTx (one-way ANOVAs; verbosity: $F(2,59) = 0.37$; $P = 0.69$; vocabulary: $F(2,59) = 0.35$; $P = 0.71$; lexical diversity: $F(2,59) = 0.013$; $P = 0.987$). (B) A histogram of interview lengths across all subjects is shown on the left, within-group averages on the right. Again, there were no statistically significant differences in interview duration (one-way ANOVA: $F(2,59) = 0.39$; $P = 0.68$). ANOVA, analysis of variance; CTTR, corrected type-to-token ratio.

First, the transcripts were preprocessed as described in the supplement (available at http://links.lww.com/PAIN/B248), after which 3 linguistic metrics were calculated for every participant: verbosity (# words), vocabulary (# unique words), and lexical diversity (corrected type-to-token ratio). These calculations were used both as internal control measures and as unique features in our machine learning (ML) classification. We also calculated idea density, psycholinguistic measures, metaphoricity, and sentiment, which are introduced here but explained in detail in the supplement (available at http://links.lww.com/PAIN/B248). *Idea density* is the number of propositions (adjectives, verbs, adverbs, prepositions, and conjunctions) divided by the total number of words, often used to measure the comprehensibility of a text[4] and the efficacy of mental processing.[21] Propositional idea density (p-density) was calculated using Computerized Propositional Idea Density Rater, version 3 (CPIDR3) and output as the percentage of ideas within an interview. *Psycholinguistic features* summarize various language components including, but not limited to, lexical, morphological, and syntactic characteristics; for our purposes, a basic set of psycholinguistic features was calculated using the Linguistic Inquiry and Word Count software (LIWC, version 2015). For each interview, we obtained 78 LIWC features spanning parts of speech, grammar, and multiple social, emotional, and cognitive word dimensions. *Metaphoricity* describes how much patients relied on figurative language in their narratives because descriptions about pain often include the use of metaphors and similes.[3,24] We followed the methods described by Gutierrez et al[17] and used the algorithm developed by Do Dinh and Gurevych[11]; the output metric was the percentage of all words in an interview labeled as metaphorical by the algorithm. Finally, *sentiment* quantifies a text's overall emotional valence and intensity and was estimated using Python's "Valence Aware Dictionary and sEntiment Reasoner" that automatically calculates positive, negative, and neutral sentiments in a text. Valence Aware Dictionary and sEntiment Reasoner produces 4 metrics—the proportion of the interview text that is positive, negative, and neutral, as well as a combination score between −1 and +1 indicative of the overall extent of negativity or positivity accounting for population norms.

Because the interviews still retained a large amount of nonmeaningful words, we removed stop words (default list from Natural Language Toolbox Kit) and lemmatized all remaining words as a normalization procedure. As the final preprocessing step, words were converted from text to number, with the number representing the alphabetized index in a selected corpus. Following recommendations by Bedi et al,[2] we used the Touchstone Applied Science Associates (TASA) collection as our corpus, consisting of thousands of text documents that represent common knowledge across the U.S. educational system (K-12). Touchstone Applied Science Associates contains 77,998 unique words that functioned as our semantic reference space. As an example, if a participant said the word "pain," it was converted to the number 50,010 as it was the 50,010th entry in TASA. Each interview was represented by a string of N numerical tokens $\{w_i\}$ to $\{w_1, w_2, \ldots, w_N\}$ that were used as the primary inputs for semantic proximity calculations.

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

Table 1

Demographic variables.

Group	Age (y)	Education (y)	Concurrent pain at 6 wks (NRS)	Concurrent positive mood at 6 wks (PANASp)	Concurrent negative mood at 6 wks (PANASn)
Continuous variables (one-way ANOVA)					
PTxNonR	44.7 ± 14.3	12.2 ± 3.5	61.6 ± 22.8	29.4 ± 11.4	14.4 ± 4.6
PTxResp	46.9 ± 11.4	12.7 ± 3.8	47.74 ± 22.8	33.4 ± 8.4	17.3 ± 6.8
NoTx	46.2 ± 11.4	13.8 ± 3.9	57.6 ± 17.5	35.9 ± 10.9	16.2 ± 10.0

Group	Gender	Race/Ethnicity	Income
Categorical variables (χ^2)			
PTxNonR	13M, 6F	[Hispanic or Latino] 3	[<$10,000] 5
		[American Indian or Alaskan Native] 0	[$10,001 to $25,000] 5
		[Asian] 0	[$25,001 to $50,000] 6
		[Black or African American] 8	[$50.001 to $75,000] 2
		[Native Hawaiian or Pacific Islander] 0	[$75,001 to $100,000] 0
		[White or Anglo-American] 8	[>$100,000] 1
		[Mixed or prefer not to answer] 0	[Prefer not to answer] 0
PTxResp	14M, 9F	[Hispanic or Latino] 6	[< $10,000] 10
		[American Indian or Alaskan Native] 0	[$10,001 to $25,000] 7
		[Asian] 0	[$25,001 to $50,000] 2
		[Black or African American] 9	[$50.001 to $75,000] 1
		[Native Hawaiian or Pacific Islander] 1	[$75,001 to $100,000] 2
		[White or Anglo-American] 6	[>$100,000] 0
		[Mixed or prefer not to answer] 1	[Prefer not to answer] 1
NoTx	10M, 10F	[Hispanic or Latino] 1	[< $10,000] 4
		[American Indian or Alaskan Native] 0	[$10,001 to $25,000] 5
		[Asian] 1	[$25,001 to $50,000] 4
		[Black or African American] 7	[$50.001 to $75,000] 3
		[Native Hawaiian or Pacific Islander] 1	[$75,001 to $100,000] 0
		[White or Anglo-American] 7	[>$100,000] 2
		[Mixed or prefer not to answer] 3	[Prefer not to answer] 2

All 3 groups were well balanced for potentially confounding demographic variables, with no differences found between responders, nonresponders, and untreated individuals in age (one-way ANOVA, $F(2,59) = 0.14$; $P = 0.87$), education (one-way ANOVA, $F(2,59) = 0.95$, $P = 0.39$), concurrent pain intensity (one-way ANOVA: $F(2,59) = 2.39$, $P = 0.10$), concurrent mood (PANASp one-way ANOVA: $F(2,59) = 2.01$, $P = 0.14$; PANASn one-way ANOVA: $F(2,58) = 0.74$, $P = 0.48$), sex (χ^2 (2) = 1.40, $P = 0.50$), race/ethnicity (χ^2 (8) = 6.80, $P = 0.56$), or income (χ^2 (12) = 12.76, $P = 0.39$).
ANOVA, analysis of variance; NRS, numeric rating scale; PANAS, Positive and Negative Affect Schedule.

2.7. Latent semantic analysis and semantic proximity

To investigate meaning more rigorously, we used the Latent Semantic Analysis (LSA[10]) on TASA, which generates a multidimensional linear representation of the semantic meaning of words based on their co-occurrence with other words in the set of texts. The frequency of word co-occurrence across different documents represents how well words are semantically related to each other.[23] Singular value decomposition was applied to break down the co-occurrence matrix X into the product of 3 matrices—U, S, and V—cropped to K = 300 dimensions[23]; the matrix U contained the semantic features for all words in TASA and functioned as our base semantic reference space. For each numericized lemma in a participant's interview, the corresponding word in U was found, resulting in an N word X 300 feature matrix for each subject (U_i) representing each patient's unique semantic space created by their interview. More details regarding singular value decomposition, the rationale for the value of K, and data organization for LSA can be found in the Supplemental Methods (available at http://links.lww.com/PAIN/B248).

Semantic proximity is a measurement of how close one word is to another in semantic space. U_i was used to calculate the semantic proximity for all words in each interview to a selection of 261 words of interest (WOIs) based on 11 hypothesis-driven themes, summarized in Figure S2, http://links.lww.com/PAIN/B248. The semantic proximity between 2 words was computed as the dot product of all 300 semantic dimensions obtained through LSA for Word A (WOI, vector from matrix U) and Word B (interview word, vector from matrix U_i); the semantic vectors were normalized to constrain the possible outcomes to be between −1 and +1. The proximity to each of the WOIs was measured for all words in an interview, and the mean was calculated to obtain an average semantic proximity value for each WOI (ie, 261 average semantic proximity values for every patient); this is represented in **Figure 1E**. All semantic analyses were run with Matlab (MathWorks, version R2016a).

2.8. Machine learning classifiers

We used 2 ML classification models to identify placebo responders vs nonresponders: Logistic Regression and Linear Support Vector Machine (SVM). All models were implemented with Python's scikit-learn package. Because no previous studies attempted to classify placebo pill responses in patients with chronic pain using language, we had no a priori hypothesis about data sparsity. Therefore, we tried 2 separate regularization approaches: L1 (lasso) and L2 (ridge) penalization, testing a total of 4 models (logistic and SVM, both with L1 and L2 penalization). Also, because of the low-sample size here, and in line with our previous study,[45] we implemented a nested leave-one-out cross-validation procedure where models were trained in an inner loop

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

(10-fold cross-validation, data from n = 41 patients) and applied to 1 left-out patient. The purpose of the inner loop was to select the number of features that best generalized to the data and to optimize the regularization parameters of each of the models (ie, L1 and L2 penalty) using a grid search. The left-out patient is thus never used to select features or optimize hyperparameters, and hence, the accuracy is unbiased.

We began with all 348 features as initial model inputs as there is, to the best of our knowledge, no preexisting literature on placebo response and language use to aid in a priori feature reduction. Within the inner loop, we split the data into folds of training and test sets; within each fold, we performed robust scaling (data were demeaned and scaled by quartiles) and feature selection (selected features ranked according to their statistical significance in an F-test using sklearn's *SelectKBest,* from a minimum of 2 up to 20 maximum features), and parameter regularization (*alpha* parameter for linear regression, *C* parameter for SVM). We used a feature selection step in addition to regularization to restrict the search space to the most significant features. This intermediate step is not strictly necessary but—given the large number of features—aids in the interpretability of the resulting model even if at a (potential) trade-off in accuracy.[38,48] Because this step was included in the inner loop, the choice of features does not bias the out-of-sample accuracy.

Once the optimal model (highest accuracy) was identified by cross-validation, it was applied to the left-out participant to classify the patient as a placebo responder (1) or nonresponder (0). This process was repeated for each subject. Models were run on all 3 blinded lists (1 real and 2 random), generating 3 sets of results before unblinding.

As a post hoc analysis, we examined whether language features could infer the magnitude of the placebo response as a continuous variable. To do this in a constrained manner, we selected the top features taken from our classification approach and built a simple linear regression model. To further refine the model and remove features of minimal contribution, we performed the regression in a forward stepwise manner given that models with several parameters are likely to overfit the data and be less generalizable to new data sets.

2.9. Examining language features—meaning, context, and relationships

Understanding language often requires a consideration of context. As an exploratory analysis, we investigated the features on a phrase level to identify the interview sections that elicited these responses. For LIWC results, we took the top 5 patients with the greatest use of features chosen by the model and found what words in their interviews had been tagged as those LIWC categories; the interview questions with the most tags were considered primary contexts for the features. For these patients, we found the 5 highest semantic proximity values or LIWC tags in their interviews and extracted the associated word and the 10 surrounding it, representing up to 5 unique phrases per person scattered throughout an interview. To better interpret meaning, we used the words associated with these 5 locations and aggregated them across all patients for a given WOI, generating 310 semantically similar words per WOI. The most frequently used words across all patients in a WOI were considered to be close approximations of meaning for our language features. Finally, we investigated the relationships between top language features by calculating their inverse covariance and comparing results between placebo response groups (PTxResp vs PTxNonR) and treatment groups (PTx vs NoTx). The inverse covariance was believed to provide an estimate of the conditional independence of the variables and was closer to a "causal" analysis as it removed common source correlations. It also allowed us to investigate underlying differences in the structure of language between the groups, as both the existence and the direction of connections between features could be informative.

2.10. Assessing the effect of treatment on language features

Although our main interest was to disentangle language use between PTxResp vs PTxNonR, our study included a no-treatment group (n = 20) to control for treatment responses. These subjects performed identical study procedures. To assess if language features were affected by simply receiving a treatment, we compared the top language features from all patients receiving treatment (PTx = PTxResp + PTxNonR) with those from patients in the no-treatment (NoTx) group. A repeated-measures analysis of variance was performed with language features as a within-subjects factor and group (NoTX and PTx) as a between-subjects factor.

2.11. Linking language features to personality traits

Previous research has demonstrated that psychological characteristics of patients with chronic pain can be mapped onto 4 personality dimensions—pain trait, character trait, aware trait, and emote trait—of which character and aware traits were found to be predictive of placebo response.[44] The character trait includes positive affect and positive personality traits, such as extraversion, openness, and conscientiousness, measured through the NEO-PI-R and Positive and Negative Affect Schedule[47]; the aware trait, in turn, is implicated in emotional regulation and several awareness capacities, as measured through the Multidimensional Assessment of Interoceptive Awareness questionnaire.[25] For a full list of all the questionnaires and subscales used for each trait, see the study by Vachon-Presseau et al. 2019.[44] In post hoc analyses, we investigated relationships between the top language features and these 4 personality dimensions. We calculated the factor scores for each of the 4 personality traits by multiplying the score of each questionnaire subscale by its corresponding binary weight (+1 for positive loadings and −1 for negative loadings). These values were then z-scored and averaged to obtain single-subject scores per trait. Pairwise correlations between each language feature and the 4 traits were calculated; significance levels were Bonferroni corrected for multiple comparisons across all 4 traits (significant cutoff is $P < 0.0125$). Associations between personality traits and language features surviving multiple comparisons were further examined through a mediation analysis (*R* package *Mediate*) to assess if language mediated the relationship between personality and placebo response (% analgesia). Finally, to see if language inferred placebo response magnitude beyond that of personality, we performed a 2-stage hierarchical regression, adding the 4 personality measures in stage 1 and the top language features in stage 2. Subsequently, we performed a model comparison (R^2 change) to assess if language explained additional unique variance in placebo analgesia.

3. Results

3.1. Participants

Of the 129 people assessed for eligibility, 125 patients with CBP enrolled into the RCT investigating placebo response propensity, outlined in **Figure 1A**. Of these, 66 completed all aspects of the

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

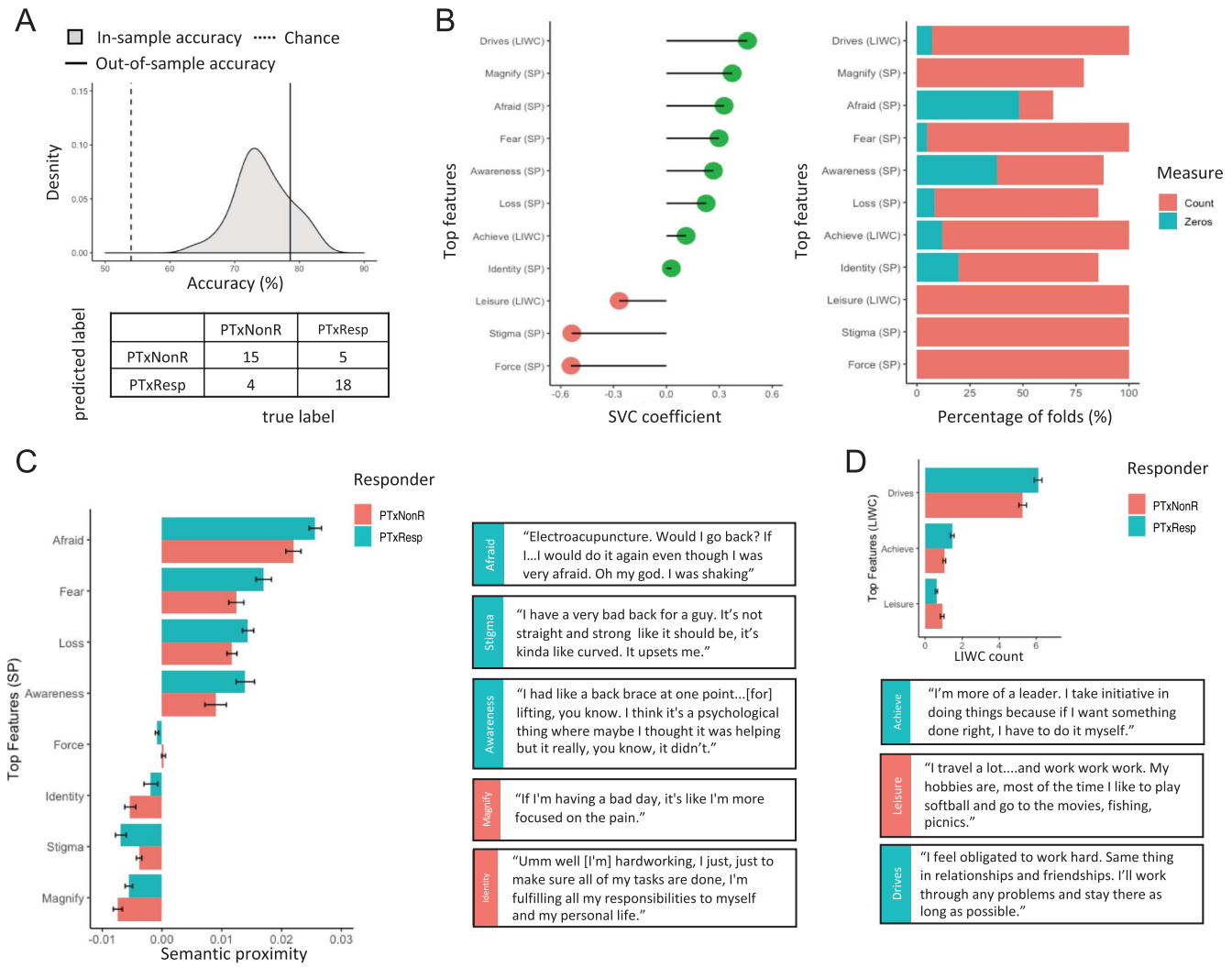


Figure 3. Machine learning model adequately differentiates placebo responders and nonresponders. (A) Model accuracy for the machine learning pipeline (SVM, L1 regularization). Density plots show in-sample training accuracy over the 42 inner cross-validation folds. Vertical line denotes out-of-sample accuracy (final model accuracy). Dotted line denotes chance level (54%). The confusion matrix demonstrates the high accuracy of the model (79% accuracy) to differentiate between placebo responders and nonresponders. (B) Model performance and model heterogeneity. Drives, afraid, and magnify show the highest positive contribution to the model, whereas force, stigma, and leisure are the highest negative contributors. Right panel shows how frequently each feature is picked up in the 42 inner folds of the nested cross-validation approach (in red), and how many times they were zeroed by L1 regularization (blue). (C) Comparisons between placebo responders (PTxResp, blue) and nonresponders (PTxNonR, red) for each of the semantic proximity top features, along with illustrative examples. Semantic proximity values closer to −1 indicate longer distances *away* from a WOI, whereas values closer to 1 indicate a shorter distance to a WOI (closer). (D) Comparison between PTxResp and PTxNonR for the LIWC, along with illustrative examples. LIWC, linguistic inquiry and word count; SVM, support vector machine; WOI, word of interest.

study, including questionnaires (Table S1, http://links.lww.com/PAIN/B248) and a semistructured open-ended exit interview at the final visit (Figure S1, http://links.lww.com/PAIN/B248). Of these patients, 4 received active treatment and were removed before analysis (as they were only for blinding purposes), making a final sample size of n = 62 patients with CBP; CONSORT diagram is given in Figure S3, http://links.lww.com/PAIN/B248. These participants included 37 men (48.8 ± 1.9 years of age) and 25 women (41.84 ± 2.7 years of age). Twenty of these participants were assigned to the no-treatment arm (NoTx); of the remaining 42 who were given placebo pills (PTx), the permutation test of pain ratings between baseline and treatment periods (**Fig. 1B**) classified 19 as placebo nonresponders (PTxNonR) and 23 as placebo responders (PTxResp).

Good compliance in pain ratings was observed at all the study moments: of the requested 2 ratings per day, an average of 1.8 ratings were observed during the baseline period and 1.7 ratings for the first and second treatment phases. For the first and second washout phases, 1.8 and 1.6 pain ratings were submitted, respectively. Importantly, no differences were observed between groups in the number of pain ratings in any of the timepoints (all Fs < 1, ps. > 0.43, Figure S4 for a breakdown of pain ratings per group and time, http://links.lww.com/PAIN/B248).

3.2. Interviews were well controlled and not influenced by obvious confounds

Interviews were audio recorded, transcribed, preprocessed, and analyzed according to the methods summarized in **Figures 1C–E**. The average length of an interview was 27.2 ± 10.3 minutes (range: 13-65 minutes). To verify that none of our participant groups differed in the general properties of their interviews (which might bias language differences seen),

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

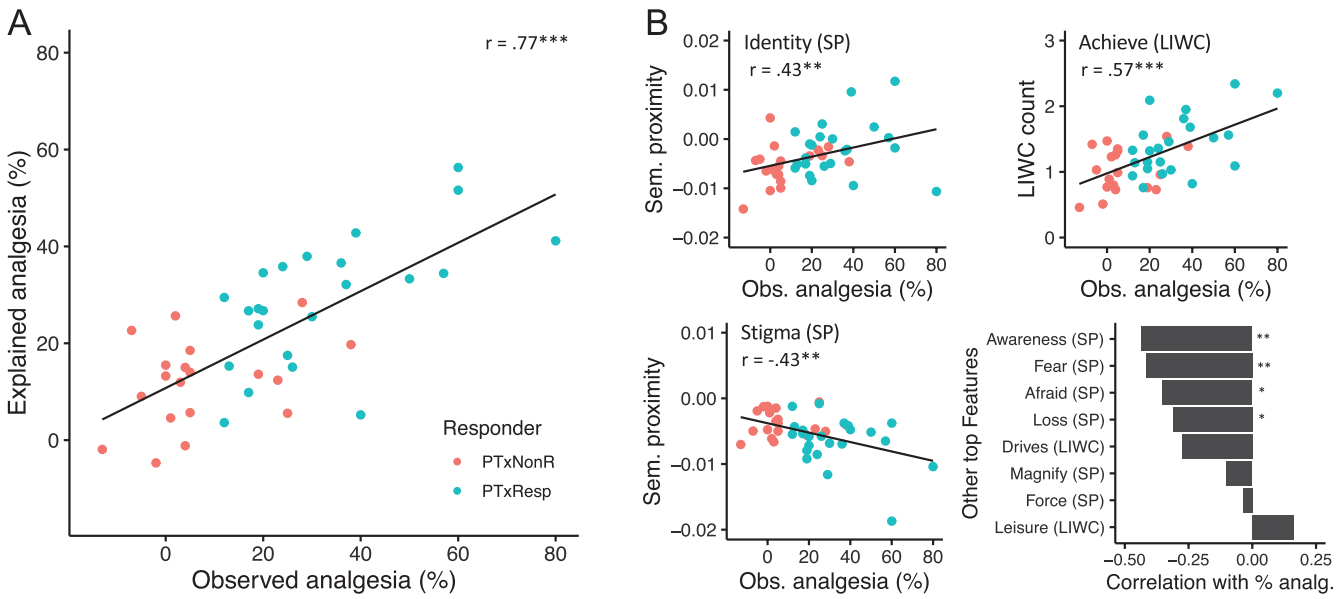


Figure 4. Top language features from the logistic model are related to the magnitude of placebo analgesia. (A) Predicted placebo analgesia of a linear regression model (after stepwise feature elimination, final features: stigma, identity, and achieve) shows a strong relationship to observed placebo analgesia. (B) Association between observed analgesia and each of the 3 features in the linear regression model. Bottom-right panel shows the correlations for the remaining top features. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

verbosity, vocabulary, and lexical diversity were calculated for each interview, averaged within group, and compared. There were no statistically significant differences in any of the 3 measures (**Fig. 2A**) nor in the lengths of the interviews between groups (**Fig. 2B**), indicating that patients were similar in their amount of talking and in basic word utilization, irrespective of presence of treatment or treatment response. None of these basic language parameters correlated with concomitant pain or mood reports, although some correlated with participants' education levels (Table S2, http://links.lww.com/PAIN/B248); we did not regress out these effects as neither education nor income differed between the 3 groups (**Table 1**). There were also no differences between groups in pain intensity, positive affect, or negative affect at the time of interview, although PTxResp had marginally lower pain scores likely due to their most recent placebo treatment effects. In addition, the 3 groups did not differ in terms of sex ratios, age, or racial/ethnic breakdown (**Table 1**).

3.3. Language feature extraction for machine learning models

We also extracted more than 300 additional features from patients' interview content, including *idea density* (a measure of thought comprehensibility and efficiency), *psycholinguistic tags* (using LIWC), *metaphoricity* (proportion of speech that was figurative), *sentiment* (extent and valence of positive and negative speech content), and *semantic proximity* (normalized cosine distance) to 261 WOIs from 11 a priori-defined themes (Figure S2, http://links.lww.com/PAIN/B248). We fed all 348 extracted language features into the ML pipeline.

3.4. Cross-validated model results

After extracting the language parameters and running the machine learning models for the 3 lists (one correct and 2 random), the results were publicly presented at a laboratory meeting, and the true list identity was revealed. The best accuracy correctly pertained to the true list. The models from the 2 random lists identified placebo responders at or below chance level (54%, shown in Figure S5, http://links.lww.com/PAIN/B248); only the results from the correctly identified list are reported in the article.

Within the inner loop, in-sample accuracies for logistic regression were 71% ± 4% and 69% ± 4 for L1 and L2 regularization, respectively. On the outer loop (ie, unbiased, out-of-sample accuracy), the accuracies of the classifiers were 74% (95% CI: 67%-81%; sensitivity: 0.73; specificity 0.75) and 67% (95% CI: 60%-74%; sensitivity: 0.78; specificity, 0.53) for L1 and L2 regularization, respectively. Using SVM substantially improved the classification, with in-sample accuracies of 74% ± 4% and 71% ± 5%, and out-of-sample accuracies of 79% (95% CI: 73%-85%, sensitivity: 0.82; specificity: 0.75) and 69% (95% CI: 61%-76%; sensitivity: 0.69; specificity: 0.69) for L1 and L2 regularization, respectively. For brevity, we report the results from the best-performing model in the article (L1-SVM, **Fig. 3A**); results from the other models can be found in the supplementary material (Figure S6, available at http://links.lww.com/PAIN/B248). Importantly, all 4 ML models identified approximately the same features with similar weights. For instance, the top 5 highest coefficients across the 4 ML models were always "fear," "awareness," "magnify," "afraid," and "drives," highlighting the consistency across classification algorithms.

Because we used a nested cross-validation approach, we effectively generated 42 models that were trained within each inner-loop iteration. It was therefore necessary to evaluate model stability to see if within each inner-loop iteration, the model was consistently selecting the same features while also assigning them similar weights. High model homogeneity indicates that these features reliably dissociate PTxResp from PTxNonR and thus are more likely to generalize to new data sets. The top features and their weights can be inspected in **Figure 3B**. Briefly, 3 semantic proximity features ("stigma," "force," and "fear") and 3 LIWC features ("drives," "achieve," and "leisure") were captured in all 42 inner-loop iterations. Other important semantic proximity features included "awareness," "loss," "magnify," "identity," and

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

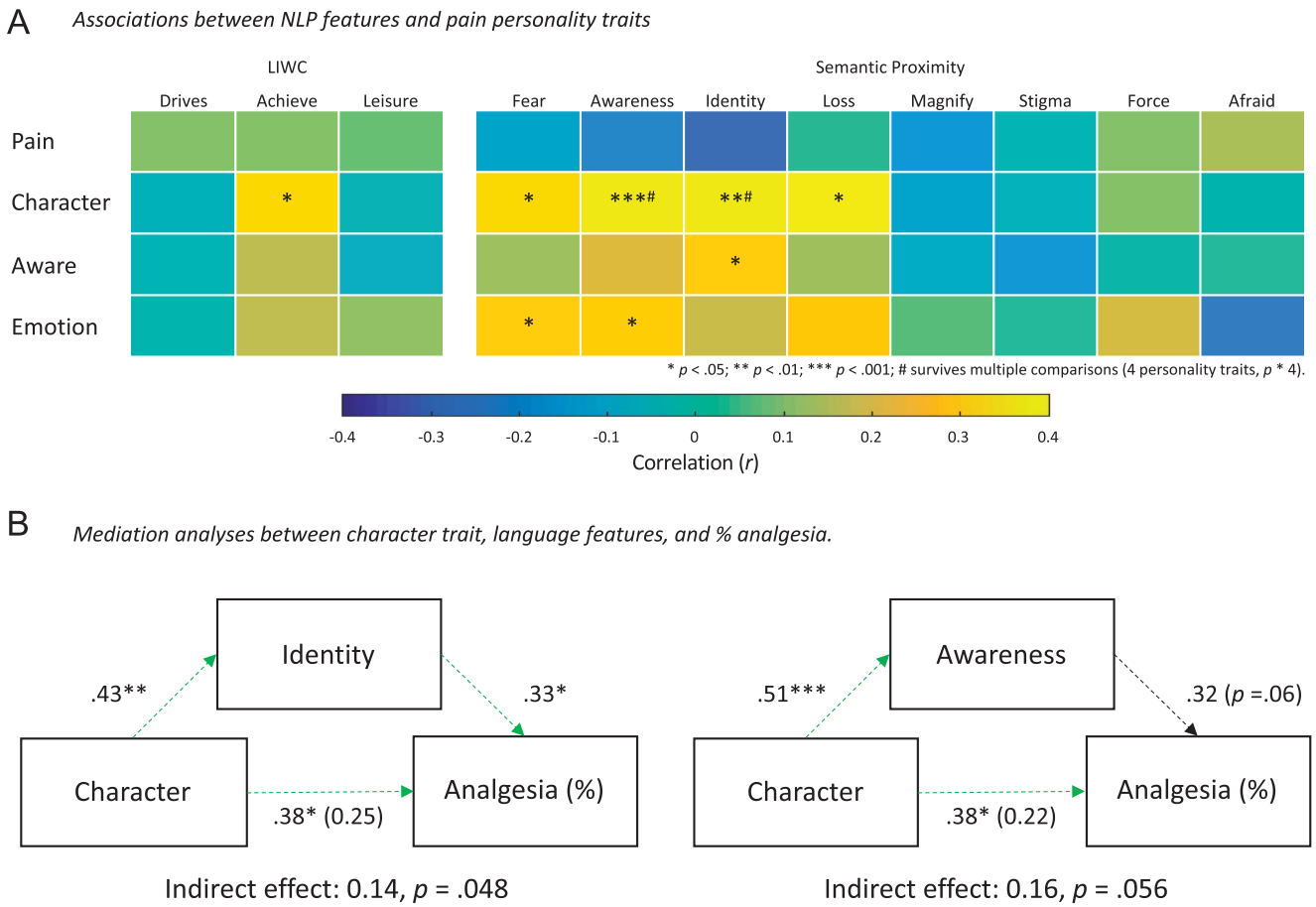


Figure 5. Associations between top language features and personality. (A) Pairwise correlations between chronic pain personality types (pain trait, character trait, aware trait, and emotion trait) and top language features. Character trait, a personality type that was previously shown to predict the magnitude of placebo analgesia, is significantly associated with multiple language features. (B) Mediation models assess whether language traits mediate the association between personality (character trait) and placebo analgesia. Both semantic proximity to identity ($P < 0.05$) and awareness ($P = 0.056$) show similar mediation effects. *$P <$ 0.05, **$P < 0.01$, ***$P < 0.001$, #surviving Bonferroni correction. Regression coefficients in mediation analyses are standardized coefficients.

"afraid" (captured in ≥ 50% of the folds). Some features were consistently chosen by the feature selection algorithm yet contributed little to the model (coefficients zeroed by L1 regularization more than half the time). These were LIWC "sexual" and semantic proximity to "empathy," "emotion," "disappoint," and "well." There were 29 additional features that were highly inconsistent (appeared, on average, in 8% of the folds with a range 2%-33%); because these features were unreliable, they were not further explored. The mean values and illustrative examples for semantic proximity and LIWC features can be seen on **Figures 3C and D**, respectively.

To assess the model's sensitivity to placebo analgesia without categorizing response, we conducted a post hoc exploratory analysis examining whether the top features from the classification model (those present in at least 50% of the folds with nonzero coefficients) could model the magnitude of placebo analgesia (% pain relief, continuous measure). A forward stepwise linear regression was conducted with all top features (LIWC: drives, achieve, and leisure; semantic proximity: stigma, force, fear, awareness, loss, magnify, and disappoint). The resulting regression model included 3 parameters: stigma (B = 0.37, $P =$ 0.004), identity (B = −0.31, $P = 0.021$), and achieve (B = −0.38, $P = 0.006$), explaining a total of 46% of the variance of the placebo response ($R^2 = 0.50$, Adj $R^2 = 0.46$, F = 12.47, $P <$ 0.001, **Fig. 4A**). These 3 language features, individually, were also strongly correlated with placebo analgesia (**Fig. 4B**). Other top features were associated with placebo analgesia (**Fig. 4B**, bottom-right panel) but were excluded from the stepwise model, most likely because of their redundancy/colinearity within the model.

3.5. Differences in language features are driven specifically by placebo responses and not by general treatment effects

As the language interviews were conducted posttreatment/end of study, the potential predictive ability of language features may be confounded by treatment effects because patients benefiting from treatment might have acquired particular language patterns from taking placebo that could be used to classify them. To control for treatment effects, our design included a no-treatment arm (NoTX group), where subjects underwent the identical experimental protocol but received no treatment. We reasoned that if the language features are similar across the treatment and no-treatment groups, these language profiles do not reflect a consequential treatment effect. To explicitly test this, a repeated-measures analysis of variance was performed, with language features as a within-subjects factor (the 9 top features included in the final model) and groups (NoTX and PTx) as a between-subjects factor. Both the main effect of Group ($F(1, 60) = 1.14$, $P = 0.29$) and the interaction Group × Features interaction ($F(10, 60) = 0.60$, $P = 0.81$) were statistically nonsignificant.

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

3.6. Language feature relationships and interpretations

We conducted a set of exploratory analyses to better understand the language features in our model. Inverse covariance between all top features revealed that PTxResp had more negative associations than PTxNonR, and that 4 of the 5 NoTx's associations were also present in PTxResp and PTxNonR graphs (Figure S7, http://links.lww.com/PAIN/B248). We also investigated where in the interviews our language features seemed based on maximum similarity or count values, with these locations providing context information. Whereas LIWC features were localized around self and event descriptions, semantic proximity features were distributed across interviews (Figure S8, http://links.lww.com/PAIN/B248). Figure S9, http://links.lww.com/PAIN/B248, shows example interview excerpts for all top language features and their corresponding interview locations. Table S3, http://links.lww.com/PAIN/B248, shows the 10 words that were the semantically closest to the WOIs in the interview and the most frequently used by patients, providing additional meaning interpretation. More details regarding these findings can be found in the supplement (available at http://links.lww.com/PAIN/B248).

3.7. Language features are associated with psychological traits and personality

To ground patients' language use to more stable traits and tendencies, we explored the relationship between language features and personality. To do so, we calculated the pairwise correlations between each feature and the 4 chronic pain personality traits from our previous study,[45] 2 of which were linked previously to placebo response: character trait and aware trait (**Fig. 5A**). Achieve ($r = 0.33$), fear ($r = 0.33$), awareness ($r = 0.52$), identity ($r = 0.45$), and loss ($r = 0.37$) showed significant correlations to character trait. Identity was also associated with the aware trait, whereas fear and awareness were associated with the emote trait; interestingly, no language features were related to the pain trait. Only the correlations of character trait with "awareness" and "identity" survived correction for multiple comparisons.

To explicitly test the association between language, personality, and analgesia, we performed a mediation analysis. Of interest was whether language could fully mediate the previously described relationship between personality and magnitude of analgesia. Because of the limited sample size, we only performed mediation analyses for the 2 strongest associations that survived multiple comparisons correction. The effect of character trait on % analgesia was mediated by the language feature "awareness." As **Figure 5B** illustrates, the regression coefficients between character trait and magnitude of analgesia, and between identity and analgesia, were significant. We tested the significance of the indirect effect using bootstrapping procedures (n = 5000), which was statistically significant ($P = 0.048$). The effect of character trait on % analgesia was partly mediated by identity, as the regression coefficient between identity and analgesia was marginally significant ($P = 0.06$). Bootstrapped significant testing revealed that the indirect effect was also marginally significant ($P = 0.06$).

We also considered whether language features could explain unique variance in placebo responses, even after accounting for personality measures. To test this, a 2-stage hierarchical multiple regression was performed, with the % analgesia as the dependent variable, personality features entered first, and "stigma," "identity," and "achieve" (3 top features) entered second. As expected, the 4 personality measures alone significantly contributed to the regression model ($F [4,36] = 3.06$, $P = 0.029$), accounting for 17% of the variance in placebo analgesia ($R^2 = 0.25$, Adj. $R^2 = 0.17$). Importantly, after entering language features, we explained an additional 31% of the variance, and this change in R^2 between stages was statistically significant ($F [3,33] = 8.33$, $P < 0.001$), showing that the predictive ability of language features go above and beyond that of personality measures. Together, personality and language features explain up to 49% of the variance in placebo analgesia ($R^2 = 0.58$, Adj. $R^2 = 0.49$).

4. Discussion

This study investigated whether the discourse of patients with CBP, analyzed using natural language processing (NLP) tools, could identify subjects who responded to a placebo. A combination of language features related to psycholinguistic content and semantic proximity successfully identified placebo responders with measurable structural differences, and the top features were able to explain substantial variance in the amount of analgesia reported. These features showed similar patterns irrespective of whether subjects received treatment or not, suggesting that components of language may be predictive of the placebo response specifically as opposed to merely due to treatment effects. Importantly, language features were significantly associated with personality traits previously shown to be predictive of the placebo response, reinforcing the plausibility of using quantitative language features to probe psychological constructs.

In line with our primary hypothesis that patient's narratives would uniquely reflect their propensity to respond to a placebo pill, we demonstrate that specific language features differentiated PTxResp from PTxNonResp posttreatment with high accuracy. These included semantic proximity to the concepts afraid, fear, awareness, force, stigma, loss, identity, and magnify, as well as the proportion of time people talked about LIWC categories achievement, drive, and leisure. Although it is difficult to directly interpret these features, some intriguing preliminary observations can be made. First, although a placebo response is often associated with positivity and optimism in healthy individuals,[15,19,22,26] there were no differences in positive sentiment values between groups nor differences in semantic proximity to positive emotions. However, closer semantic proximity with words associated with negative emotions and experiences (eg, afraid, fear, and loss) significantly differentiated PTxResp and PTxNonR. PTxResp also showed a positive conditional relationship between fear awareness and negative relationship between loss stigma and afraid stigma. These findings may reflect the idea that placebo responders are more aware of their negative emotions, an interpretation that is in line with our previous findings.[45] Importantly, closer semantic proximity to negative words does not necessitate that PTxResp experienced increased negative emotions; instead, it reflects that responders talked more about concepts related to these words. It is possible that responders may be more willing to talk about negative emotions and experiences than nonresponders, a characteristic which might relate to personality traits such as openness, which has been shown to be important for a placebo response.[50] Another compelling finding is that the 2 groups showed differences in semantic proximity to awareness. Previous studies have linked placebo analgesia to interoceptive awareness[44,45] and, more generally, to body consciousness and increased embodiment tendencies.[19] Thus, the result that closer semantic proximity to awareness identified PTxResp fits well with previous evidence. In

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

addition, PTxResp had a greater proportion of their language pertaining to leisure and less pertaining to achievements or drive than PTxNonR, differences which may be related to personality features known to influence the placebo response in other populations, including fun seeking, goal seeking, self-efficacy, and self-esteem.[19]

Other top features chosen by the model are less interpretable, including semantic distance to the concepts magnify, force, stigma, and identity. When considering the words that were not only the closest or most similar to these WOIs in the interview but also the most frequently used by patients, a better approximation for meaning may be considered. For example, force is associated with words that can be generated by an individual ("pull," "push," and "lift") or the consequences of such forces ("rest"). When combining this information with the finding that force occurs the most often during patients' descriptions of their pain experience and pain influencers, we may be picking up signals about how their body and associated pain has been impacted by certain kinds of movements or events. Magnify has semantic links with words related to size ("small") and perspective ("focus" and "describe"); its peaks occurred most often during patients' discussion of pain influencers, coping, and alternative therapies. Thus, this WOI likely captures patients' perceptions of what may make their pain worse or better. Words highly associated with identity include "experience," "personal," "emotional," "attitude," and "self" among others, which makes sense given its occurrence during questions related to self-descriptions and responses about how pain has affected their sense of self. Stigma is the most challenging to interpret. It seems to have the highest semantic signal in questions related to how pain affects mood and patients' experiences with the medical system. But because of polysemy, it is difficult to untangle its multiple disparate meanings that are all represented in its vector and our WOIs (eg, a mark of shame or disgrace, an indication of illness, and the organ of a flower). The words most closely associated with stigma and most commonly used by patients include "male," "female," and "sex," seemingly biased towards the biological definition instead of the social one. Given that it is very unlikely our patients talked specifically about flower reproduction in their interviews (although they may have spoken about how chronic pain impacted their sexual functioning or relationships), we surmise that stigma is picking up both aspects related to biological functioning as well as social aspects of stigmatization, which may lead to comparisons in status or treatment ("even" or "same"). However, all of these interpretations remain speculative at best. Additional research as well as a larger sample sizes would be needed to investigate and understand these features further. Yet, even in the absence of definitive interpretation, the proof-of-concept that language can be used as a tool to objectively dissociate placebo response types was largely successful.

In addition to identifying placebo responders with high accuracy, we also found that language features can explain the magnitude of pain relief after placebo ingestion. Measuring the analgesic potential of a placebo can provide valuable information about the therapeutic value of placebo pills, and this method can potentially be used to determine the utility of placebo in clinical settings. Furthermore, it provides an efficient way to control for placebo effects in double-blind, randomized controlled trials, discounting the placebo effect from the effect of active drugs. This has the added value of controlling for the effect size of placebo, rather than eliminating all placebo responders from the study.

We chose an exit interview to study language because talking about chronic pain could be an emotionally charged experience and we did not want any potentially therapeutic or distressing effects bleeding into future placebo/nocebo responses (suggestibility and emotional valence attributed to treatment are key predictors of placebo response[19]); this unfortunately limited our ability to directly test the predictive capacity of our language features. Considering that the interview was conducted post-treatment, it was necessary to examine if these language patterns could have potentially predicted placebo response had they been collected pretreatment or, instead, reflected consequences of general treatment effects that caused the differences seen between groups. We reasoned that if language patterns were impacted by treatment alone, then we should see differences in language features between all patients who received treatment (PTx) and those who did not (NoTX), as well as group differences in conditional relationships between features. Importantly, this was not the case, suggesting that the effects in language seen were not a consequence of response but rather a component of or predisposition to response.

In further support of this previous statement, we found significant associations between language features (awareness and identity) and chronic pain personality traits that have been shown to be predictive of placebo response before treatment administration. This suggests that these language features potentially reflect stable manifestations of personalities of patients with chronic pain.[44] Because the personality measures were collected before treatment, it implies that these language patterns were present before the start of the treatment itself. Indeed, the evidence that language can be grounded within chronic pain personality traits reflects the nontransient nature of these features. Naturally, because of the correlational nature of these findings and the time discrepancy (pretreatmen vs posttreatment), caution is warranted in this interpretation, and causal evidence supporting this claim is required. We will pursue this in future studies.

This study demonstrates that language-related metrics can be used to quantify placebo response and chronic pain relief in a clinical trial. Importantly, although we previously demonstrated that questionnaires[45] and chronic pain personality traits[44] alone are able to predict who will respond to placebo and explain a sizable amount of variance of the analgesia, the results shown here improve on our previous modelling attempts. In our former study, personality measures alone were able to predict placebo responders at 72% accuracy (vs 79% here), and explained 30% of the magnitude in analgesia (vs 46% here). This suggests that outside of simply providing more explanatory power to self-reported scores, language features may pick up additional information that is lost while using standardized tests. This is further reinforced by our mediation analyses, which shows that half of the predicted analgesia effect from personality is mediated by select language features.

The finding that adding language features to a model with personality leads to a substantial and significant increase in explained variance suggests that language is tapping into a unique and subtle aspect of placebo response. This is perhaps not surprising considering the high success of NLP in classifying other kinds of patient groups.[5,6,8,27] This improvement in accuracy might signify that this measure is both more ecological and also more sensitive because it provides a vast amount of data to predict from; our constrained analyses left us with 348 features, but many more could have been added. For instance, a number of studies have shown promising relationships between pain intensity and acoustic features of voice, including breathiness,[28] rate,[35] and loudness[41]; however, acoustic features were not calculated in this study because the environment was not well

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

controlled for sound quality. Likewise, expanding our list of WOIs, including different corpora and semantic spaces, and using more NLP tools to calculate additional semantic and syntactic features would also be warranted in future iterations of this research. Given this, it is reasonable to expect the accuracy of placebo prediction model using multiple kinds of speech features can be further improved.

In conclusion, we have demonstrated that quantitatively analyzing speech of a patient with pain in an RCT using NLP and ML techniques can successfully identify placebo responders and estimate the magnitude of placebo analgesia. Importantly, these language features are related to stable personality traits found in patients with chronic pain and are specific to placebo-related effects instead of general treatment effects. Together, these findings provide additional insight into the mechanisms at play in placebo analgesia and provide novel methodology to study, predict, and objectively quantify placebo effects in clinical studies.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Acknowledgments

The authors want to thank all Apkarian laboratory members who contributed to this study with their time and resources (in particular Alex Baria who scrambled and reorganized the data for us). The authors would also like to thank all patients who participated in this study for their overall commitment during the trial and their honesty and candidness during the interviews. Finally, the authors want to thank Elkin Dario Gutierrez and Rachel Ostrand from IBM Research for their help in implementing the metaphoricity and CPIDR analyses, respectively, and Elif Eyigoz and Carla Agurto from IBM Research for their discussions and feedback about the machine learning analyses and inverse covariance, respectively. This work was funded by National Center for Complementary and Integrative Health AT007987. EVP was funded through Canadian Institutes of Health Research (CIHR) and Fonds de Recherche Santé Québec (FRQS). 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/B248.

Supplemental video content

A video abstract associated with this article can be found at http://links.lww.com/PAIN/B249.

Article history:
Received 27 July 2020
Received in revised form 16 October 2020
Accepted 9 November 2020
Available online 5 January 2021

References

[1] Agurto C, Cecchi GA, Norel R, Ostrand R, Kirkpatrick M, Baggott MJ, Wardle MC, Wit Hde, Bedi G. Detection of acute 3,4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing. Neuropsychopharmacology 2020;45:823–32.

[2] Bedi G, Cecchi GA, Slezak DF, Carrillo F, Sigman M, de Wit H. A window into the intoxicated mind? Speech as an index of psychoactive drug effects. Neuropsychopharmacol 2014;39:2340–8.

[3] Biro D. The language of pain: finding words, compassion, and relief. 1st ed New York: W.W. Norton, 2010.

[4] Brown C, Snodgrass T, Kemper SJ, Herman R, Covington MA. Automatic measurement of propositional idea density from part-of-speech tagging. Behav Res Methods 2008;40:540–5.

[5] Cecchi G, Corcoran C. O2.3. Automated analysis of recent-onset and prodromal schizophrenia. Schizophr Bull 2018;44:S76.

[6] Clarke N, Foltz P, Garrard P. How to do things with (thousands of) words: computational approaches to discourse analysis in Alzheimer's disease. Cortex 2020;129:446–63.

[7] Colloca L. The placebo effect in pain therapies. Annu Rev Pharmacol Toxicol 2019;59:191–211.

[8] Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, Bearden CE, Cecchi GA. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiat 2018;17:67–75.

[9] Dahlhamer J, Lucas J, Zelaya C, Nahin R, Mackey S, DeBar L, Kerns R, Von Korff M, Porter L, Helmick C. Prevalence of chronic pain and high-impact chronic pain among adults—United States, 2016. MMWR Morb Mortal Wkly Rep 2018;67:1001–6.

[10] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391–407.

[11] Do Dinh EL, Gurevych I. Token-level metaphor detection using neural networks. Proceedings of the Fourth Workshop on Metaphor in NLP, San Diego, CA: Association for Computational Linguistics, 2016. pp. 28–33. doi:10.18653/v1/W16-1104.

[12] Finniss DG, Kaptchuk TJ, Miller F, Benedetti F. Biological, clinical, and ethical advances of placebo effects. Lancet 2010;375:686–95.

[13] Flaten MA, Aslaksen PM, Lyby PS, Bjørkedal E. The relation of emotions to placebo responses. Philos Trans R Soc Lond B Biol Sci 2011;366:1818–27.

[14] Fuentes J, Armijo-Olivo S, Funabashi M, Miciak M, Dick B, Warren S, Rashiq S, Magee DJ, Gross DP. Enhanced therapeutic alliance modulates pain intensity and muscle pain sensitivity in patients with chronic low back pain: an experimental controlled study. Phys Ther 2014;94:477–89.

[15] Geers AL, Weiland PE, Kosbab K, Landry SJ, Helfer SG. Goal activation, expectations, and the placebo effect. J Pers Soc Psychol 2005;89:143–59.

[16] Gupta U, Chatterjee N. Personality traits identification using rough sets based machine learning. 2013 International Symposium on Computational and Business Intelligence. New Delhi: IEEE, 2013. p. 182–5. doi:10.1109/ISCBI.2013.44.

[17] Gutierrez ED, Cecchi G, Corcoran C, Corlett P. Using Automated Metaphor Identification to Aid in Detection and Prediction of First-Episode Schizophrenia. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. pp. 2923–2930. doi:10.18653/v1/D17-1316.

[18] Hashmi JA, Baliki MN, Huang L, Parks EL, Chanda ML, Schnitzer T, Apkarian AV. Lidocaine patch (5%) is No more potent than placebo in treating chronic back pain when tested in a randomised double blind placebo controlled brain imaging study. Mol Pain 2012;8:29.

[19] Horing B, Weimer K, Muth ER, Enck P. Prediction of placebo responses: a systematic review of the literature. Front Psychol 2014;5:1079. doi:10.3389/fpsyg.2014.01079.

[20] Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW. Language style matching predicts relationship initiation and stability. Psychol Sci 2011;22:39–44.

[21] Kemper S, Sumner A. The structure of verbal abilities in young and older adults. Psychol Aging 2001;16:312–22.

[22] Kern A, Kramm C, Witt CM, Barth J. The influence of personality traits on the placebo/nocebo response. J Psychosom Res 2020;128:109866.

[23] Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104:211–40.

[24] Lascaratou C. The Language of Pain: Expression or description? Amsterdam: John Benjamins Publishing Company, 2007. doi:10.1075/celcr.9.

[25] Mehling WE, Price C, Daubenmier JJ, Acree M, Bartmess E, Stewart A. The multidimensional assessment of interoceptive awareness (MAIA). PLoS One 2012;7:e48230.

[26] Morton DL, Watson A, El-Deredy W, Jones AKP. Reproducibility of placebo analgesia: effect of dispositional optimism. PAIN 2009;146:194–8.

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.

[27] Orimaye SO, Wong JS-M, Golden KJ, Wong CP, Soyiri IN. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinformatics 2017;18:34.
[28] Oshrat Y, Bloch A, Lerner A, Cohen A, Avigal M, Zeilig G. Speech prosody as a biosignal for physical pain detection. Proc. Speech Prosody 2016: 420–24.
[29] Padfield M, Procter I. The effect of interviewer's gender on the interviewing process: a comparative enquiry. Sociology 1996;30: 355–66.
[30] Peciña M, Azhar H, Love TM, Lu T, Fredrickson BL, Stohler CS, Zubieta JK. Personality trait predictors of placebo analgesia and neurobiological correlates. Neuropsychopharmacol 2013;38:639–46.
[31] Pennebaker JW, King LA. Linguistic styles: language use as an individual difference. J Pers Soc Psychol 1999;77:1296–312.
[32] Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. Annu Rev Psychol 2003;54: 547–77.
[33] Pennebaker JW, Stone LD. Words of wisdom: language use over the life span. J Pers Soc Psychol 2003;85:291–301.
[34] Pollner M. The effects of interviewer gender in mental health interviews. J Nerv Ment Dis 1998;186:369–73.
[35] Roy N, Volinn E, Merrill RM, Chapman CR. Speech motor control and chronic back pain: a preliminary investigation. Pain Med 2009;10: 164–71.
[36] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman MEP, Ungar LH. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 2013;8:e73791.
[37] Sonnenschein AR, Hofmann SG, Ziegelmayer T, Lutz W. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. Cogn Behav Ther 2018;47:315–27.
[38] Spisak T, Kincses B, Schlitt F, Zunhammer M, Schmidt-Wilcke T, Kincses ZT, Bingel U. Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. Nat Commun 2020;11:187.
[39] Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 2010;29:24–54.
[40] Tétreault 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.
[41] Thiam P, Kessler V, Amirian M, Bellmann P, Layher G, Zhang Y, Velana M, Gruss S, Walter S, Traue HC, Kim J, Schork D, Andre E, Neumann H, Schwenker F. Multi-modal pain intensity recognition based on the SenseEmotion database. IEEE Trans Affect Comput 2019:1.
[42] Toivonen R, Kivelä M, Saramäki J, Viinikainen M, Vanhatalo M, Sams M. Networks of emotion concepts. PLoS One 2012;7:e28883.
[43] 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.
[44] 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.
[45] 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.
[46] Wager TD, Atlas LY. The neuroscience of placebo effects: connecting context, learning and health. Nat Rev Neurosci 2015;16:403–18.
[47] Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988;54:1063–70.
[48] Xiang ZJ, Wang Y, Ramadge PJ. Screening tests for lasso problems. IEEE Trans Pattern Anal Mach Intell 2017;39:1008–27.
[49] Yarkoni T. Personality in 100,000 Words: a large-scale analysis of personality and word use among bloggers. J Res Personal 2010;44:363–73.
[50] Yu R, Gollub RL, Vangel M, Kaptchuk T, Smoller JW, Kong J. Placebo analgesia and reward processing: integrating genetics, personality, and intrinsic brain activity. Hum Brain Mapp 2014;35:4583–93.

Copyright © 2021 by the International Association for the Study of Pain. Unauthorized reproduction of this article is prohibited.