Group-regularized individual prediction: theory and application to pain

Martin A Lindquist, Anjali Krishnan, Marina López-Solà, Marieke Jepma, Choong-Wan Woo, Leonie Koban, Mathieu Roy, Lauren Y Atlas, Liane Schmidt, Luke J Chang, et al.

▶ **To cite this version:**

Martin A Lindquist, Anjali Krishnan, Marina López-Solà, Marieke Jepma, Choong-Wan Woo, et al.. Group-regularized individual prediction: theory and application to pain. NeuroImage, 2017, 145 (Pt B), pp.274-287. 10.1016/j.neuroimage.2015.10.074 . hal-04590519

HAL Id: hal-04590519

https://hal.sorbonne-universite.fr/hal-04590519

Submitted on 28 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2018 January 15.

Published in final edited form as:

Neuroimage. 2017 January 15; 145(Pt B): 274–287. doi:10.1016/j.neuroimage.2015.10.074.

Group-regularized individual prediction: Theory and application to pain

Martin A. Lindquist[1], **Anjali Krishnan**[2], **Marina Lopez-Sola**[2], **Marieke Jepma**[2], **Choong-Wan Woo**[2], **Leonie Koban**[2], **Mathieu Roy**[3], **Lauren Y. Atlas**[4], **Luke J. Chang**[2], **Liz Losin**[2,5], **Hedwig Eisenbarth**[2], **Yoni K. Ashar**[2], **Zeb Delk**[2], and **Tor D. Wager**[2]

[1]Johns Hopkins University

[2]University of Colorado Boulder

[3]Concordia University

[4]National Center for Complementary and Integrative Health, National Institutes of Health

[5]University of Miami

Abstract

Multivariate pattern analysis (MVPA) has become an important tool for identifying brain representations of psychological processes and clinical outcomes using fMRI and related methods. Such methods can be used to predict or 'decode' psychological states in individual subjects. Single-subject MVPA approaches, however, are limited by the amount and quality of individual-subject data. In spite of higher spatial resolution, predictive accuracy from single-subject data often does not exceed what can be accomplished using coarser, group-level maps, because single-subject patterns are trained on limited amounts of often-noisy data. Here, we present a method that combines population-level priors, in the form of biomarker patterns developed on prior samples, with single-subject MVPA maps to improve single-subject prediction. Theoretical results and simulations motivate a weighting based on the relative variances of biomarker-based prediction—based on population-level predictive maps from prior groups—and individual-subject, cross-validated prediction. Empirical results predicting pain using brain activity on a trial-by-trial basis (single-trial prediction) across 6 studies (N = 180 participants) confirm the theoretical predictions. Regularization based on a population-level biomarker—in this case, the Neurologic Pain Signature (NPS)—improved single-subject prediction accuracy compared with idiographic maps based on the individuals' data alone. The regularization scheme that we propose, which we term group-regularized individual prediction (GRIP), can be applied broadly to within-person MVPA-based prediction. We also show how GRIP can be used to evaluate data quality and provide benchmarks for the appropriateness of population-level maps like the NPS for a given individual or study.

Please address correspondence to: Tor D. Wager, Department of Psychology and Neuroscience, University of Colorado, Boulder, 345 UCB, Boulder, CO 80309, tor.wager@colorado.edu, Telephone: (303) 895-8739.

Matlab code for all analyses is available at: http://www.columbia.edu/cu/psychology/tor/

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

Machine learning; statistical learning; pain; MVPA; Empirical Bayes; prediction; fMRI; mega-analysis; meta-analysis; shrinkage

Introduction

Tremendous progress has been made in fMRI research over the past 10 years. Much of the benefit has resulted from multivariate pattern analysis (MVPA) based studies of mental representations, which have enhanced the ability to identify brain patterns that are predictive of behavioral and psychological outcomes (Chang et al., In Press; Davis and Poldrack, 2013; Haxby et al., 2014; Haxby et al., 2001; Kay et al., 2008; Poldrack et al., 2009; Wager et al., 2013; Woo et al., 2015). In standard brain mapping analyses, many regions of the brain might respond to a given task. However, for a pattern of brain activity to be considered useful as a *representation* of a psychological or behavioral state, it must be predictive of (sensitive and specific to) that state.

Recent studies have identified provisional representations for many kinds of psychological states, including perception of low level visual features (Kamitani and Tong, 2005) and higher-order object properties (Haxby et al., 2001), knowledge of semantic categories (Huth et al., 2012; Mitchell et al., 2008), memory (Kuhl et al., 2011; Rissman et al., 2010; Xue et al., 2010), affective states such as pain (Brodersen et al., 2012; Cecchi et al., 2012; Marquand et al., 2010; Wager et al., 2013), and emotion (Baucom et al., 2012; Chang et al., In Press; Kassam et al., 2013), and identification of individuals with clinical disorders (Arbabshirani et al., 2013; Craddock et al., 2009; Doehrmann et al., 2013; Fu et al., 2008; Siegle et al., 2006; Whelan et al., 2014). Once representations of specific percepts (e.g., objects) or experiences (e.g., emotion) are identified, studies can examine how these representations are shaped by contextual, psychological, and neurobiological processes—e.g., how object representations are maintained in working memory during a delay (Harrison and Tong, 2009), how items are recollected during memory recall and compete with other memories (Kuhl et al., 2011), or how pain representations are modified by cognitive reappraisal (Woo et al., 2015). Identifying patterns of fMRI activity that can serve as proxies for representations requires multivariate analyses that are predictive of outcomes in individual subjects. In this paper, we develop a method for improving such single subject, MVPA-based predictions.

Most single subject predictive analyses utilize only data from one participant in developing the predictive model (e.g., (Horikawa et al., 2013). The theory behind this approach is that brain representations are idiographic, i.e., different individual subjects have different multivariate brain patterns that predict outcomes. For example, the pattern of fMRI activity within V1 that predicts line orientation may be different for different individuals (Freeman et al., 2011; Kamitani and Tong, 2005), and only patterns at broader spatial scales may be conserved across individuals (Heeger and Ress, 2002; Norman et al., 2006). If brain topography is truly idiographic and varies dramatically across individuals, individualized training to derive the best predictive multivariate brain pattern is likely the optimal strategy.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

However, often, there is information at multiple spatial scales, including much information conserved across individuals (Chang et al., In Press; Kassam et al., 2013; Poldrack et al., 2009; Rissman et al., 2010; Shinkareva et al., 2008; Wager et al., 2013; Woo et al., 2014). In addition, the quantity and quality of fMRI data is limited in single subject datasets, and often high-quality single subject prediction requires hours of scanning for each individual over multiple days (Gonzalez-Castillo et al., 2012; Nishimoto et al., 2011). Often, perhaps surprisingly, models that are trained to predict out-of-sample individuals perform as well or better than models trained on individual subject data (Chang et al., In Press; Poldrack et al., 2009; Shinkareva et al., 2008; Wager et al., 2013) when the spatial topography of predictive information is shared across individuals. In such cases, using normative group maps based on other individuals may help to regularize single subject predictive patterns using information conserved across subjects, constraining the single subject solution in ways that improve prediction accuracy and prevent overfitting.

This paper develops a principled scheme for combining normative group maps based on previously defined predictive patterns (i.e. signatures) and single subject idiographic maps. In addition to improving prediction accuracy, this procedure regularizes individual subject maps towards prior expectations, therefore improving the quality of single-subject predictive maps, and allowing for a principled updating of normative population-based maps as data is accumulated. This weighting can be expressed in both frequentist and Bayesian frameworks, which are shown to be mathematically equivalent. Marquand et al (2014) addressed a similar problem, by recasting the decoding problem in a multi-task learning framework, allowing them to extract information from the data by sharing information between subjects. This was found to be extremely beneficial when only a small number of trials were available for each subject.

The method we develop here, which we term group-regularized individual prediction (GRIP), combines group and idiographic maps in proportion to their respective variances, in accordance with theory on empirical Bayes estimation. It can be applied prospectively to individual subjects' data to improve prediction accuracy and stabilize individual-subject predictive maps. Thus, one main use is in improving single-subject MVPA-based prediction accuracy. In addition, it can be used to provide quality control estimates and benchmarks for a given individual or study paradigm. The quality of idiographic predictions can be used to benchmark data quality for individual persons, and the accuracy of prediction using a population-level map can provide benchmarks on the appropriateness of the map for a given population, sample, or study paradigm. Such cross-study metrics are valuable as fMRI data are increasingly used in multi-site and translational settings.

The GRIP method can be applied to any domain and is agnostic with respect to the training algorithm used. However, in this paper, we evaluate its utility in predicting pain intensity ratings. Pain is an interesting application domain for three reasons. First, it is associated with enormous cognitive, social, and economic costs (IOM, 2011), but its neurological bases are not yet well understood (Tracey, 2011). Developing brain models capable of predicting pain intensity and dissociating different types of neurological contributions to pain is a high-priority. Second, pain is currently assessed primarily by means of self-report, a behavioral measure of subjective experience that is compromised in many vulnerable populations (e.g.,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the very old or very young, persons with cognitive impairment, and those who are minimally conscious) and influenced by a number of complex sociocultural factors. Brain-based predictive models could complement self-report by providing measures of neurophysiological systems that contribute to pain, and ultimately identify sub-types of pain and sub-types of patients based directly on brain information. And third, population-level maps predictive of pain intensity are available (Wager et al., 2013), providing priors to use in regularizing individual-subject predictions. Several groups have published innovative work on single subject prediction (Brodersen et al., 2012; Cecchi et al., 2012; Marquand et al., 2010). Complementing these approaches, we have developed a normative population-based pattern that classifies stimuli differing moderately in pain intensity with over 90% accuracy, across multiple sites and scanners and in new, out-of-sample individuals (Wager et al., 2013). Here, we combine information from this population-normed signature pattern—called the Neurologic Pain Signature (NPS)—with idiographic MVPA maps to improve the accuracy of predicting pain intensity from brain activity.

We begin by developing the statistical theory underlying empirical Bayes regularization and the GRIP model. We then present brief theoretical simulations that characterize the conditions under which weighting towards individuals versus group maps is optimal. Then we apply the method to combined data from six studies of experimental thermal pain (N = 180), comparing the accuracy of (a) cross-validated idiographic predictive maps, (b) a population-level map, the NPS, and (c) the GRIP combination of the NPS prior and idiographic maps (see Fig. 1 for an overview). Predictions are made about single trials, i.e. individual periods of thermal stimulation lasting 1.85 to 15 seconds, using time series-appropriate cross-validation methods. The results show that the GRIP estimator outperforms both the population-level NPS map and the idiographic, single-subject prediction map.

Method

Theory

Suppose we have a set of observations from m trials of a certain stimulus applied to a single subject, which we denote $(\boldsymbol{x}_j, y_j)$ for $j = 1, \ldots m$. Here, $\boldsymbol{x}_j$ is a vector of features of length V, and y_j is a scalar outcome variable. In our example, we assume that each trial consists of a thermal stimulus. Thus, $\boldsymbol{x}_j$ is a summary of the brain response, and y_j is the reported pain corresponding to that trial.

Now, suppose we seek to use these observations to create a predictive model from which we can estimate pain report from brain activation for the subject in question. Using standard machine learning techniques (the approach is agnostic to the specific type of technique, though we assume that it is linear in the continuation) we can find a set of idiographic brain weights $\hat{\boldsymbol{w}}_I$ that can be used to predict the outcome. Using these weights, we can predict the pain response corresponding to features $\boldsymbol{x}^*$ as follows:

$$\hat{Y}_I = \hat{\boldsymbol{w}}_I^T \boldsymbol{x}^*. \quad [1]$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the continuation we will refer to $\hat{\boldsymbol{w}}_I$ as the *idiographic map*. Given infinite amounts of training data, it may be difficult to improve on $\hat{Y}_I$ without improving the model basis fundamentally; if the model is specified correctly, $\hat{Y}_I$ will be unbiased (accurate) for subject *I*. With limited data, however, $\hat{Y}_I$ may be quite noisy (low precision), and provide a poor approximation of Y_I.

Further, suppose that prior research has provided us with a population-level based biomarker $\hat{\boldsymbol{w}}_P$, which we can similarly use to predict the outcome. Using these group-level weights, we can predict the pain response corresponding to features $\boldsymbol{x}^*$ as follows:

$$\hat{Y}_P = \hat{\boldsymbol{w}}_P^T \boldsymbol{x}^*. \quad [2]$$

In the continuation we will refer to $\boldsymbol{w}_P$ as the *population-level map*. To the degree that subject *I* differs from the population mean, $\hat{Y}_P$ provides a biased (inaccurate) estimate of Y_I. However, because $\hat{Y}_P$ has generally been estimated using significantly more data, i.e., many subjects, it may be substantially more precise. The smaller the differences between subject *I* and the population mean, the more using w_P^T will improve on $\hat{Y}_I$.

Now we have predictions for the same trials based on both a single-subject idiographic map and a population-level map, each with different strengths and limitations. The goal is to find a way to combine these two measures in a principled manner that increases the prediction accuracy compared to using each of the maps in isolation.

S*hrinkage estimators* (Efron and Morris, 1975; James and Stein, 1961) are adaptive weighting schemes that have been shown to improve upon many traditional statistical estimators — in terms of mean squared error (MSE) — by shrinking these estimators towards some fixed constant value. Shrinkage is implicit in Bayesian inference, penalized likelihood inference, and multi-level models (Lindquist and Gelman, 2009), and is directly related to the empirical Bayes estimators commonly used in neuroimaging data analysis (Friston and Penny, 2003; Friston et al., 2002; Mejia et al., 2014; Shou et al., 2014; Su et al., 2009). In all Bayesian analyses, posterior estimates are 'shrunk' towards prior expectations. In empirical Bayes shrinkage, posterior estimates are shrunk towards a prior derived by estimating a population-level (group) distribution.

In this work, we illustrate how shrinkage estimators can be used to improve upon the prediction accuracy obtained using idiographic maps developed on single subject data. We do so by shrinking the prediction towards that obtained using population-level maps. Our approach combines such group-level priors, in the form of biomarker patterns developed on prior samples, with individual MVPA predictive weights. This combination can be used to improve single-subject prediction accuracy compared with idiographic training based on the individual's data alone. The regularization scheme that we propose, which we term group-regularized individual prediction (GRIP), can be applied broadly to within-person MVPA-based prediction.

In general, the GRIP estimator corresponds to the shrinkage estimator defined as follows:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\hat{Y}_G = \lambda \hat{Y}_I + (1-\lambda)\hat{Y}_P \quad [3]$$

where λ — the shrinkage factor — can take any value in the range[0,1]; see Fig. 1 for an illustration. When λ = 0, the subject-specific data are considered completely unreliable and the estimator is reduced to the group result (reducing to 'between-participant' prediction). In contrast, when λ = 1, the subject-specific data are deemed perfectly reliable and the estimate is not shrunk towards the group value at all. In practice, of course, data are seldom perfectly reliable, which makes applying some shrinkage towards the population mean a good alternative in many situations.

A major question is how to estimate the appropriate value for the shrinkage factor λ associated with the GRIP estimator. In this paper we explore two different approaches towards choosing the appropriate shrinkage factor. The first uses a standard empirical Bayes approach based on the ratio of the within- and between-subject variances, and the second is based upon the cross-validated prediction accuracy of the idiographic and group training, respectively. Below, we describe each approach in turn.

Empirical Bayes Approach—In the *Empirical Bayes (EB) approach* we base the shrinkage factor on the ratio of the within- and between-subject variances. To quantify these values let us assume that the true pain report (Y) for subject *i* on the *j*th trial, where *j* = 1, … *M*, can be modeled as follows:

$$\begin{aligned} Y_{ij} &\sim N(\mu_i, \sigma^2_{W,i}) \\ \mu_i &\sim N(\mu, \sigma^2_B) \end{aligned} \quad [4]$$

In words this implies that the reported pain in trial *j* follows a normal distribution with subject-specific mean μ_i and variance $\sigma^2_{W,i}$. The subject mean is, in turn, a draw from a normal distribution with mean μ, the population mean, and variance σ^2_B.

In our problem, the term μ_i represents the true reported pain in subject *i* to a stimulus, while the term μ represents the corresponding mean reported pain in the population. Similarly, the variation of the reported pain across trials around the subject mean is given by $\sigma^2_{W,i}$, and the variation of the true subject-specific report around the population mean is given by σ^2_B.

Here we assume that μ_i can be estimated by $\hat{Y}_I$, as shown in Eq. 1, and μ can be estimated by $\hat{Y}_P$, as shown in Eq. 2. The within-subject variance can be estimated by computing the variance of the observed pain reports and those predicted by the idiographic map, which we denote $\hat{\sigma}^2_{W,i}$. The variance of the observed pain reports to those predicted by the population-level map, in turn, gives us an estimate of the total variance $\sigma^2_{W,i} + \sigma^2_B$, which we denote $\hat{s}^2$. Using these two results we can estimate the between-subject variance as $\hat{\sigma}^2_B = max\left\{0, \hat{s}^2 - \hat{\sigma}^2_{W,i}\right\}$, with the maximum taken to ensure non-negative estimates.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

With these values in hand we can now compute the shrinkage factor for a given subject as follows:

$$\lambda_i = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_{W,i}^2}. \quad [5]$$

Here, if the between-subject variance is large relative to the within-subject variance, then $\lambda = \lambda_i$ will be close to 1 and the idiographic map is weighted higher. If, in contrast, the within-subject variance dominates, then λ will be close to 0 and the idiographic map is weighted lower. This approach is equivalent to the Best Linear Unbiased Predictor (BLUP) of μ_{ij}, as well as the mode of the posterior distribution in a normal-normal model where μ is the posterior mean. Hence, the proposed method can be understood both within a frequentist and Bayesian framework. The term λ is estimated for each subject based on the precision of the estimates of $\hat{\sigma}_{W,i}^2$ (e.g., the residual variance from single-subject prediction). With many variables in the predictive model, minimally biased estimates can be obtained based on cross-validation. In neuroimaging experiments, for example, there are typically many predictors (voxels) and $\hat{\sigma}_{W,i}^2$ is estimated based on leave-one-run out cross-validation.

Using the estimate of the shrinkage factor, and assuming a linear classifier as those shown in Eqs. 1 and 2, we can obtain the subject-specific GRIP map as follows:

$$\hat{\boldsymbol{w}}_G = \lambda \hat{\boldsymbol{w}}_I + (1-\lambda)\hat{\boldsymbol{w}}_P \quad [6]$$

This weight can now be applied to independent data from that particular subject to predict their pain ratings associated with specific observed features $\boldsymbol{x}^*$ in test data.

Cross-validated Approach—In the *cross-validated approach* we base the shrinkage factor on the prediction accuracy obtained using the idiographic and population-level maps, respectively. This requires performing a two-level nested cross-validation procedure. To elaborate, we begin by splitting the data into K-folds (e.g., K runs). We then perform an outer cross-validation loop, which consists of K steps. For each step, K-1 folds are used as training data and a single fold is left out and used as test data. A second, inner cross-validation loop is then performed on the training data, providing us with pain predictions for each trial in the training data. Next, we use the population-level map to predict the same pain reports. We then find the best linear combination of the idiographic and population-level predictions in terms of optimizing the correlation with the true reported pain in the training data (which is assumed to be known).

To illustrate, let $\hat{\boldsymbol{Y}}_I$ and $\hat{\boldsymbol{Y}}_P$ be vectors of the idiographic and population-level predictions in the training data, while $\mathbf{Y}$ is the vector of true responses. We now set λ to be the value that maximizes:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\max_v \left\{ corr \left(v\hat{\boldsymbol{Y}}_I + (1-v)\hat{\boldsymbol{Y}}_P, \boldsymbol{Y} \right) \right\}. \quad [7]$$

This procedure is repeated for each step in the outer loop, with each fold rotating as the training data, giving us a different shrinkage factor for each of the K folds. Finally, we average the shrinkage factor across folds, and use this value to compute the GRIP weight as described in Eq. 6. This weight can now be applied to independent data from that subject to predict the pain associated with the observed features $\boldsymbol{x}^*$.

Simulation

In this section we provide two brief theoretical simulations to illustrate the properties of the Empirical Bayes approach. In the first simulation we assume that the magnitude, variance within participants, and variance between participants are all fixed (μ=1, σ_{within} = 1.2, and $\sigma_{between}$ = 0.3, respectively, based on reasonable values from prior studies), and allow the number of trials per participant to vary. Fig. 2A shows the single-trial prediction accuracy as a function of the number of trials per participant. Here it is clear that group-level predictions are unaffected by the number of trials, while idiographic predictions improve as the number of training trials per participant increases. As a result, the GRIP prediction, which is a combination of both the group and idiographic maps, will also increase with more training trials per participant. Eventually, (in results not shown here) the idiographic predictions will converge to the GRIP results as the number of trials increase.

In the second simulation we assume the magnitude, number of trials per participant, and variance between participants are all fixed (μ=1, n=50, $\sigma_{between}$ = 0.3, respectively), while the variance within participants is allowed to vary. Fig. 2B shows single trial prediction accuracy as a function of λ, which is the GRIP weighting factor for combining group and idiographic weight maps. When the within participant variance is high (i.e., the yellow curve), the optimal value of λ tends to be closer to 0, weighting the group estimate more than the idiographic estimate. In contrast, when the within-subject variance is lower (i.e., the red curve), λ will peak closer to 1, thus weighting the idiographic data higher.

Materials and Procedures

Participants—The analysis included a total of 209 healthy participants (before exclusion criteria were applied) from 7 independent studies, with sample sizes ranging from N = 17 to N = 50 per study. Descriptive statistics on the age, sex, and other features of each study sample are provided in Table 1. Participants were recruited from New York City and Boulder/Denver Metro Areas. The institutional review board of Columbia University and the University of Colorado Boulder approved all the studies, and all participants provided written informed consent. Preliminary eligibility of participants was determined through an online questionnaire, a pain safety screening form, and a functional Magnetic Resonance Imaging (fMRI) safety screening form.

We applied several exclusion criteria for analysis purposes. Participants with psychiatric, physiological or pain disorders, neurological conditions, and MRI contraindications were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

excluded prior to enrollment. For these analyses, participants from Study 1 (N = 26) were excluded because these data were used to compute the population-level biomarker (i.e., Neurologic Pain Signature; NPS). They are included in this report (e.g., Table 1) so that readers can compare stimulation and pain levels with the other studies. In addition, to have enough data for within-person cross-validation, we required participants to have at least 23 trials with low variance inflation factors (< 2.5; see below), and non-missing painful heat rating and stimulation intensity data. Based on these criteria, an additional 3 participants were excluded, resulting in a total of 180 participants for the final analyses.

Procedures—In all studies, participants received a series of contact-heat stimuli and rated their experienced pain following each stimulus. The number of trials, stimulation sites, inter-trial intervals, rating scales, and stimulus intensities and durations varied across studies, but were comparable; these variables are summarized in Tables 2 and 3. Each study also comprised a specific psychological manipulation, such as placebo treatment, which will be or has been reported elsewhere (Table 3). In the present paper, we focus on cross-validated prediction of pain report across all trials, irrespective of the study-specific psychological and physical manipulations that influenced pain.

Thermal stimulation—In each study we delivered thermal stimulation to multiple skin sites using a TSA-II Neurosensory Analyzer (Medoc Ltd., Chapel Hill, NC) with a 16 mm Peltier thermode endplate (Study 7: 32 mm). On every trial, after the offset of stimulation, participants rated the magnitude of the warmth or pain they had felt during the trial on a visual analog scale. Other thermal stimulation parameters varied across studies, with stimulation temperatures ranging from 40.8°C to 50°C and stimulation durations from 1.85 to 12.5 sec. Most studies applied thermal stimulation to the forearm. See Table 2 for stimulation intensity levels, mean temperature for each intensity level, and details of the rating scales. See Table 3 for stimulation duration, duration of inter-stimulus interval, number and location of stimulation sites, and number of trials per subject.

fMRI Analysis

Preprocessing—Structural T1-weighted images were co-registered to the mean functional image for each subject using the iterative mutual information-based algorithm implemented in SPM (Ashburner and Friston, 2005), and were then normalized to MNI space using SPM. SPM versions varied across studies (Studies 1 and 6 used SPM5; all other studies used SPM8 (http://www.fil.ion.ucl.ac.uk/spm/)). Following SPM normalization, Studies 1 and 6 included an additional step of normalization to the group mean using a genetic algorithm-based normalization (Atlas et al., 2010; Atlas et al., 2014; Wager and Nichols, 2003).

In each functional dataset, we removed initial volumes to allow for image intensity stabilization (see Table 4 for number removed in each study). Prior to processing functional images, we removed volumes with signal values that were outliers within the time series (i.e., “spikes”). To identify outliers, we first computed both the mean and the standard deviation of intensity values across each slice, for each image. Mahalanobis distances for the matrix of (concatenated) slice-wise mean and standard deviation values by functional volumes (over time) were computed. Any values with a significant χ^2 value (corrected for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

multiple comparisons based on the more stringent of either false discovery rate or Bonferroni methods) were considered outliers. In practice, less than 1% of images were deemed outliers. The outputs of this procedure were later included as nuisance covariates in the first level models. Next, functional images were corrected for differences in the acquisition timing of each slice and were motion-corrected (realigned) using SPM. The functional images were warped to SPM's normative atlas (warping parameters estimated from co-registered, high-resolution structural images), interpolated to 2×2×2 mm^3 voxels, and smoothed with an 8 mm FWHM Gaussian kernel.

Single trial analysis (Except Study 3 and Study 6)—For each study we employed the single trial, or "single-epoch", design and analysis approach to model the data. Quantification of single-trial response magnitudes was done by constructing a GLM design matrix with separate regressors for each trial, as in the "beta series" approach (Mumford et al., 2012; Rissman et al., 2004). First, boxcar regressors, convolved with the canonical hemodynamic response function (HRF), were constructed to model cue, pain, and rating periods in each study. Then, we included a regressor for each trial, as well as several types of nuisance covariates. Because each trial consisted of relatively few volumes, trial estimates could be strongly affected by acquisition artifacts that occur during that trial (e.g. sudden motion, scanner pulse artifacts, etc.). Therefore, trial-by-trial variance inflation factors (VIFs; a measure of design-induced uncertainty due, in this case, to colinearity with nuisance regressors) were calculated, and any trials with VIFs that exceeded 2.5 were excluded from the analyses. For Study 1, we also excluded global outliers (trials that exceeded three standard deviations (SDs) above the mean), and employed a principal components based denoising step during preprocessing to minimize artifacts. This approach generated single trial estimates that reflect the amplitude of the fitted HRF on each trial and refer to the magnitude of anticipatory and pain-period activity for each trial in each voxel.

Single trial analysis (Only Study 3 and Study 6)—For Studies 3 and 6, single trial analyses were based on fitting a set of three basis functions, rather than the standard HRF used in the other studies. This flexible strategy allowed the shape of the modeled hemodynamic response function (HRF) to vary across trials and voxels. This procedure differed from that used in other studies because (a) it maintains consistency with the procedures used in the original publication on Study 6 (Atlas et al., 2010), and (b) it provides an opportunity to examine predictive performance using a flexible basis set. For both Study 3 and Study 6, the pain period basis set consisted of three curves shifted in time and was customized for thermal pain responses based on previous studies (Atlas et al., 2010; Lindquist et al., 2009). To estimate cue-evoked responses for Study 6, the pain anticipation period was modeled using a boxcar epoch convolved with a canonical HRF. This epoch was truncated at 8 s to ensure that fitted anticipatory responses were not affected by noxious stimulus-evoked activity. As with the other studies, we included nuisance covariates and excluded trials with VIFs > 2.5. In Study 6 we also excluded trials that were global outliers (those that exceeded 3 SDs above the mean). We reconstructed the fitted basis functions from the flexible single trial approach to compute the area under the curve (AUC) for each trial and in each voxel. We used these trial-by-trial AUC values as estimates of trial-level anticipatory or pain-period activity.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Predictive model—For each subject, the input data consisted of their single trial data. Throughout the analyses, we used a machine-learning-based regression technique, least absolute shrinkage and selection operator-regularized principal components regression (LASSO-PCR; (Wager et al., 2011), to predict pain reports from the fMRI activity. The LASSO-PCR procedure is a combination of several established techniques chosen to work well for fMRI data. It first uses principal components analysis (PCA) to reduce the dimensionality of the data first, and thereafter applies LASSO regression (Hastie et al., 2005) to the component scores. The latter step provides a principled way of selecting a subset of distributed components and weights that together best explain the training data. We do not expect the results to vary strongly as a function of the algorithm used, as LASSO-PCR has produced virtually identical results to support vector regression in our previous work (Wager et al., 2013).

The idiographic estimates were obtained by performing a cross-validated HV block (Racine, 2000) to estimate pain reports in test data set. We begin by ordering the trials in temporal order. For a given trial, we selected the v observations immediately before and after it, thus creating a test data set consisting of $2v+1$ trials. Thereafter, we removed the h trials collected immediately before and after the test set as a buffer, and used the remaining elements to form the training set. Hence, the value of v controls the size of the test set and the value of h controls the dependence between the training and test set, following the idea that the correlation between trials decreases as they become further apart from each other. The procedure was performed in a manner so that each trial was represented in the test data one time. In our application, we chose both h and v to take the value 4. This approach is consistent for stationary observations, as the probability of selecting the model with the best predictive properties converges to 1 as the total number of observations becomes large. Using this approach cross-validated idiographic estimates were obtained for each trial. The NPS estimates were obtained by directly applying the NPS map to the data.

A total of three combination GRIP estimates were obtained for comparison purposes. The first estimate, which we denote the ‘Oracle’ estimate, was obtained by finding the linear combination of NPS and idiographic predictions that maximizes the correlation with the true pain reports (see Eq. 7). Note that this approach assumes that the true pain reports are known. So, while the approach is not feasible in practice, in our setting it provides a benchmark of how much improvement shrinkage-based approaches can potentially give. The second estimate was obtained using the ‘CV-approach’ described above. Here, for each step in an outer cross-validation loop, created using the same cross-validated HV blocks procedure as described above, we performed an inner 10-fold cross-validation step. Within each step of the inner loop we computed an ‘oracle’ weighting as described above. This allowed us to compute the optimal weight for each partition, which we subsequently applied to test data. The third estimate was computed using the ‘EB-approach’, in which weights are computed according to Eq. 5.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results

Idiographic brain maps

Idiographic (single-subject) predictive weight maps were subjected to group analyses to identify consistently predictive regions. Each weight map was obtained using the model fit on the whole data. Bootstrap tests were used to provide p-values for voxel weights in order to threshold predictive weights for display and interpretation. First, we constructed 5,000 bootstrap samples (with replacement) consisting of paired brain and outcome data and ran LASSO-PCR on each. Two-tailed, uncorrected P-values were calculated for each voxel based on the proportion of weights below or above zero. Fig. 3 shows average predictive maps for all subjects in all test samples (Studies 2–7; Study 1 was not included because it was the training dataset for the Neurologic Pain Signature prior map) and for each individual study, thresholded at .010444, which corresponds to the $q < .05$ False Discovery Rate corrected p-value across all 180 subjects. In addition, it shows one example subject (from Study 2, with idiographic predictive accuracy $r = .5$) both before and after shrinkage (using $\lambda = 0.65$) to the study mean. The average maps shows that the basis for single-subject predictions is reproducible across individuals, particularly within a core set of brain regions associated with encoding noxious stimuli and correlated with pain reports in a number of previous studies. As this paper focuses primarily on methodology, we do not interpret these patterns beyond noting their inter-subject consistency, though we describe them briefly below.

Increased activity predicted increased pain (yellow in Fig. 3) in a number of regions directly and indirectly targeted by nociceptive afferents. Those that are direct targets include ventrolateral thalamus, medial thalamus, periaqueductal gray and surrounding midbrain nuclei, hypothalamus, and medullary activity consistent with the parabrachial complex. Other regions with positive weights are often considered part of ‘pain processing’ systems and receive direct input from regions targeted by primary nociceptive afferents, including bilateral dorsal posterior insula (dpINS), primary somatosensory cortex (S1), secondary somatosensory cortex (S2), anterior dorsal and ventral insula, the dorsal anterior cingulate and mid-cingulate, and large areas of the cerebellum. In addition, some regions are predictive of pain but do not, to our knowledge, have direct projections from nociceptive systems, including portions of the dorsomedial prefrontal cortex and parahippocampal cortex extending into occipital regions. These systems may contribute to pain independent of nociceptive input.

In addition, decreased activity predicted increased pain (blue in Fig. 3) in a number of brain regions. These included reproducible contributions from the ventromedial prefrontal cortex, precuneus, medial orbitofrontal cortex, ventrolateral and dorsolateral prefrontal cortices, lateral parietal and sensorimotor cortices, more anterior portions of the hippocampus/parahippocampal cortex, superior temporal cortices, and lateral occipital regions.

There appear to be some visible differences in the average weights across studies, though a full analysis of inter-study variability requires much more detailed analyses and is beyond the scope of the present paper. The paradigms were similar, but differed in the nature of the psychological and behavioral manipulations performed in addition to variations in noxious

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

stimulus intensity. Of particular note, however, are the maps for Study 7, which was the only study to use heat pain on the calf rather than the hand for all trials. Notably, this study was the only to show reproducible positive weights in medial somatosensory regions, which is where somatosensory input from the foot and leg is mapped within S1. However, as our goal here is primarily to develop the GRIP method, we do not assess inter-individual and inter-study variability further. For present purposes, many of the core 'pain-processing' regions (e.g., dpINS and anterior cingulate) showed consistent, positive predictive weights in all six studies. This consistency demonstrates that the predictive maps are reproducible and interpretable in the sense that the regions involved are associated with nociceptive pathways as corroborated by neuroanatomical and functional measures in invasive animal and human studies.

Prediction accuracy across studies

All four methods were substantially above chance, on average and for each individual study, in predicting single-trial pain ratings. Our primary accuracy measure was correlations between predicted and observed pain reports across single trials, averaged across participants. These values were $r = 0.32$, 0.34, and 0.38 for NPS, idiographic, and GRIP methods (all $p < .0001$). These correlations are calculated across single trials; if correlations are computed on the averages over several repeated trials (e.g., averaged pain over 8 trials, with correlations calculated across intensity levels), the correlation values will be substantially higher. As the classification analyses below show, these effect sizes can yield highly accurate classification of high vs. low pain in individual participants, and in fact are consistent with the effect sizes reported in previous studies (e.g., Wager et al. 2013), which report correlations in the range of $r = 0.5 – 0.8$ when on the order of 4–11 trials per condition are averaged.

Comparing these accuracy values across shrinkage methods provides a way of evaluating their performance. Fig. 4A shows a violin plot (i.e., a box plot with a rotated kernel density plot appearing on each side) of the single-trial prediction-outcome correlations across the entire dataset (N = 180) for each approach computed using the NPS map, an idiographic map, and three combination maps (using the cross-validated, empirical Bayes and oracle approaches). Not unexpectedly, the oracle approach outperforms the other four methods, as it is able to find the optimal post hoc combination of individual-subject and group predictive map weights; thus, this serves as our benchmark. However, of the methods that can be applied *a priori*, the EB approach performs best. It outperforms the CV-approach ($t(179) = 3.84$, $p < 0.0005$), which in turn outperforms the idiographic estimate ($t(179) = 2.06$, $p < 0.05$). Finally, the idiographic estimate performs significantly better than the NPS estimate ($t(179) = 13.13$, $p < 0.0001$). Fig. 4B shows similar results for the NPS, Idiographic, CV, and EB maps separated by study. Though, the relative quality of the NPS and Idiographic maps vary across studies, the CV and EB approaches consistently outperform both. In addition, the difference in the predictive performance of each approach varies significantly across studies. For example, Study 7 for which painful stimuli were applied to the leg (and not the forearm) and were significantly shorter in duration (1s as opposed to 10s durations), showed significantly lower prediction accuracies for all approaches.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Fig. 5 shows the average shrinkage factors across subjects for the EB and CV approaches, both combined and separated by study. Both produce similar weights for the idiographic versus population-level maps, with weights around 0.6 for nearly all studies, indicating a moderate bias towards using idiographic data but strong contributions from both sources. Values of 0.5 would indicate equal contributions from the idiographic and population-level maps. One would expect the weights to favor the population-level maps with smaller amounts of individual data, higher individual noise, or higher model dimensionality (more independent voxels). Conversely, one would expect the weights to favor the idiographic maps when the population-level model is a less accurate description of the individual's basis for pain, caused either by differences in the type or location of pain or perhaps individual differences in cortical organization. Notably, Study 7 also shows a stronger bias towards idiographic maps, indicating that the NPS is a weaker predictor of pain in this study. This is sensible because, as noted above, Study 7 was the only one to involve leg pain.

In order to provide an interpretable metric for predictive accuracy beyond prediction-outcome correlations, we calculated the equivalent forced-choice accuracy in classifying high vs. low pain given the effect sizes observed in our studies. In other words, we sought to determine how often one would classify a more painful condition as being more painful than a less painful one, given the signature response (i.e., the expression of the signature pattern) associated with each. We started with the signature response-pain report correlation across single trials for each approach, averaged across participants and studies. These values, r = 0.32, 0.34, and 0.38 for NPS, idiographic, and GRIP, respectively, constitute a measure of effect size that can be converted into classification accuracy for any given number of trials. First, we converted the average Pearson's r values to Cohen's d, using the formula: $\mathrm{d}=2\mathrm{r}/\sqrt{1-r^2}$. Then, assuming equal numbers of trials (n) for both conditions being compared, an effect magnitude of 1 standard deviation increase in pain, and equal variances in the high and low pain conditions, we obtained a z value for the sampling distribution of the difference between the average pattern response in high and low conditions: $z=d\sqrt{n/2}$. The accuracy is based on the long-term average proportion of the correct responses, acc = $\Phi(z)$, where Φ represents the cumulative distribution function of the normal distribution.

Fig. 6 shows the resulting forced-choice accuracy curves, which provide an estimate of accuracy (y-axis) as a function of number of trials (x-axis) that may provide a useful and interpretable benchmark for future studies. The curves show the classification accuracy for the original NPS, Idiographic prediction, and GRIP Empirical Bayes combination. Accuracy depends on the effect size, or how strongly the predictive map is related to single-trial outcome measures, and the number of trials averaged. The plot shows that to achieve 90% accuracy with the NPS, one needs at least 7 trials per condition, or 14 trials. To achieve 90% accuracy with the Idiographic maps, one needs at least 6 trials per condition, or 12 trials. Finally, to achieve 90% accuracy with the GRIP combination, one needs only 5 trials per condition, or 10 trials. This is approximately 30% fewer trials to achieve this criterion. Conversely, greater accuracy is possible for a given amount of data, depending on the effect size. While all models converge on high accuracy with larger numbers of trials (>20 per condition, or 40 trials) in Fig. 6 due to the large effect size, with smaller effect sizes the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

benefits of the GRIP method will be appreciable even for situations with larger numbers of trials.

Assessing data quality

Fig. 7A demonstrates a theoretical approach for assessing certain aspects of data quality and model fit. The predicted outcomes for participants based on the idiographic weight maps can be plotted against their predicted outcomes based on the NPS weight map. Four quadrants are determined based on the predictions on each axis. Participants in the top right quadrant have highly predictive population-level and idiographic weight maps. These “canonical good” participants have high data quality and normative brain function that is reflective of the population’s brain function. In contrast, participants in the lower left quadrant have low data quality, resulting in a bad model fit with both population and idiographic maps. The low data quality could be due to bad behavioral ratings, noisy brain data, or data processing errors. “Idiosyncratic” participants in the lower right quadrant have highly predictive idiographic weight maps, but the population-level weight map is not very predictive. The successful idiographic prediction indicates good data quality, while failed population-level prediction may reflect non-normative brain function and organization. Finally, the participants in the upper left quadrant show good prediction by the population-level weight map, but poor prediction from their idiographic maps. Such individuals are likely to have good data quality with minimal processing errors, but there is either insufficient training data for successful idiographic prediction or variation in prediction-outcome relationships across cross-validation folds (runs). Individuals in this quadrant illustrate the value of using population-level data even when individualized predictive maps are available. Combining population-level and idiographic prediction may thus be informative for determining data quality.

These are plots of predictive accuracy, which can serve to identify cases in which data quality or model fit is poor. For example, a subject with low predictive accuracy for both idiographic and group-based maps is likely to have one of several problems, including (a) poor data quality/artifacts, (b) model mis-fit or mis-specification at the time series level, (c) rating biases, such that ratings do not reflect underlying experience/processes, among other possibilities. We cannot distinguish among these alternatives, but the plots are useful in checking whether there are problems that could be addressed more systematically with additional analyses and quality control checks. In addition, other types of patterns in the predictive plots can place other kinds of constraints on inferences about data quality and model fit. For example, high idiographic but low group-based predictive accuracy rules out all three of the problems (a–c) above, and suggest that the individual has a non-normative brain basis for the outcome (pain experience).

Figure 7B shows an example of such a data quality plot for participants from six out of the seven studies (Study 1 was not included because it was the training dataset for the Neurologic Pain Signature prior map). The four theoretical quadrants are identified using a threshold of $r > 0.2$ for the predicted outcome from idiographic and NPS weight maps. The upper right quadrant includes participants with highly predictive outcomes from both idiographic and NPS weight maps (proportion of participants per study—Study 2: 63%;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Study 3: 71%; Study 4: 66%; Study 5: 72%; Study 6: 71%; Study 7: 35%). The lower right quadrant includes participants with highly predictive outcomes from the idiographic maps, but non-predictive outcomes from the NPS (Study 2: 13%; Study 3: 21%; Study 4: 12%; Study 5: 7%; Study 6: 0%; Study 7: 15%). The upper left quadrant includes participants with insufficient training data for idiographic prediction, but good population-level data (Study 2: 10%; Study 3: 4%; Study 4: 14%; Study 5: 14%; Study 6: 18%; Study 7: 4%). The lower left quadrant includes participants with non-predictive outcomes for both idiographic and NPS weight maps, and these might be considered for exclusion (Study 2: 13%; Study 3: 4%; Study 4: 8%; Study 5: 7%; Study 6: 11%; Study 7: 46%). These results suggest that Study 7 differs from the NPS prior more than other studies. Study 7 was the only study in which the stimulus was applied to the leg instead of the arm, and the stimulation time was significantly shorter in duration (1s as opposed to ~10s duration). Idiographic predictions were also lower for Study 7, which is consistent with the reduced time-on-task with brief stimulus durations. However, the reduced power in single-trial analyses is balanced by the fact that more trials can be obtained with brief stimulation. Thus, it is correct that data quality per trial is lower in Study 7, though overall assessment of quality should also consider how much data can be collected and what the study goals are (e.g., single-trial prediction vs. prediction or mapping of trial averages).

Discussion

This paper proposes a simple and robust method for combining normative population-level predictive maps with idiographic maps in order to optimize brain-based prediction and classification of individual-subject outcomes. We apply this method to prediction of pain intensity, using trial-by-trial brain maps to predict trial-by-trial ratings of pain intensity either idiographically (within-subject) or using idiographic predictive maps regularized (shrunk) towards the NPS, a population-level predictive map trained on independent data (Wager et al., 2013). In six independent test datasets (N = 17–50 each), using a GRIP estimator with the shrinkage factor chosen based on the ratio of the within- and between-subject variances (an Empirical Bayes approach) gives a significant increase in prediction accuracy compared to using single-subject training data alone. Though we tested the model on pain, the procedure we applied could be useful for brain-based prediction and classification in any domain—e.g., identification of face- or place-selective regions (Haxby et al., 2001) or other perceptual mapping (Formisano et al., 2008; Kamitani and Tong, 2005), predicting subsequent memory (Johnson et al., 2009; Rissman et al., 2010; Xue et al., 2010), and semantic classification (Huth et al., 2012; Mitchell et al., 2008), among others. In addition, because our method effectively regularizes predictive maps, it could be used to improve single-subject estimates of brain representations (patterns) as well as prediction accuracy. However, as neuroimaging data can exhibit substantial inter-subject variability in functional anatomy (Frost and Goebel, 2012), it may not always be optimal to directly link the predictive weights to a common average. In these situations it may be required to map subjects into a common reference space such that the activity patterns from different subjects can be meaningfully coupled (Haxby et al., 2011).

Below, we discuss (a) our recommended method for regularization; (b) principles governing when our findings are likely to generalize and how to apply them in new domains; (c)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

additional uses of the method in establishing data quality and quality control; and (d) applications to single subject single trial settings.

Regularization method

We compared two regularization methods to both the use of the population-based NPS map alone or idiographic map alone. The former constitutes the use of a group-level map to make predictions for out-of-sample individuals, and is approximated by "between-participant" cross-validation (Chang et al., In Press; Poldrack et al., 2009; Shinkareva et al., 2008; Wager et al., 2013). The latter uses within-participant training and cross-validation, which is the "classic" way MVPA has been used in most neuroimaging studies (Baucom et al., 2012; Brodersen et al., 2012; Cecchi et al., 2012; Haxby et al., 2001; Kamitani and Tong, 2005; Kassam et al., 2013; Rissman et al., 2010; Xue et al., 2010). The two hybrid methods both estimated an optimal weighting of the population-based and idiographic predictive maps for a given individual participant. They were: (a) the use of Empirical Bayes weighting based on the ratio of between-participant and within-participant accuracy, and (b) the use of nested cross-validation to estimate the optimal weighting for each individual. Both methods require some cross-validation to determine the within-participant accuracy, but (b) requires two-level nested cross-validation to estimate both the within-participant accuracy and the shrinkage factor. The Empirical Bayes approach is theoretically preferred to the degree that the Normal model that underlies it is a good approximation of the underlying data distribution. The nested cross-validation approach makes fewer assumptions, but is likely to be less precise with limited data.

Here, the Empirical Bayes approach outperformed the nested cross-validation approach, and we recommend it as a default approach towards estimating the shrinkage factor. However, in some applications, such as prospective prediction or predictive map estimation with limited training data, it may be desirable to choose a shrinkage factor *a priori*, obviating the need for any cross-validated variance estimation. Here, the shrinkage factor was relatively consistent across studies (though see below), and a default value of 0.6 (i.e., 60% individual, 40% population) is a reasonable starting point. The optimal value will of course depend strongly on the appropriateness of the population-level map for the test participant, which we discuss in more detail below.

Generalization and applications to new domains

We tested each of the approaches—population-based, idiographic, and the two hybrids—using six fMRI studies of thermal pain, which differed according to a number of design and acquisition parameters. Specifically, the studies differ in the intensity of the painful stimuli (different temperatures), the duration of the stimuli, body stimulation sites, the number of pain trials, inter-trial intervals, and, importantly, the presence of different psychological manipulations (e.g., placebo interventions and expectancy manipulations) that aimed at changing the experience of pain by potentially utilizing different brain networks and processes than the ones considered when deriving the NPS normative population-level map.

These differences may influence the results to varying degree for the different studies. For example, even if for all studies, the combined empirical Bayes approach appears to be the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

one showing the highest cross-validated predictive accuracy, the predictive performance of each of these approaches, and the differences between them (NPS-population based, idiographic or GRIP approach) varies significantly across studies. For example, the study for which painful stimuli were applied to the leg (and not the forearm) and were significantly shorter in duration (1s as opposed to 10s durations), showed significantly lower prediction accuracies for all three of the predictive approaches (NPS, idiographic and GRIP). However, GRIP clearly outperformed the other two approaches. This specific study showed the worst NPS-population prediction performance, suggesting that the specific characteristics of the brain response to pain in this study were less well represented by the NPS population norm than for any of the other studies. This is in agreement with the fact that the NPS was developed using significantly longer stimuli, which were applied to the forearm (instead of the leg, which evokes less pain intensity perceptions). Had the population-based signature been developed using a more similar painful stimulus and bodily location, the population-based predictor would have no doubt shown much higher performance for this particular study. The observation that the idiographic predictor also performed significantly worse than the average of the other studies suggests that indeed this study has the least reliable within-subject pain-evoked brain signal, when compared with the other. This points towards the neurobiology of pain processing suggesting that, in accordance with previous evidence, stimulation applied to the leg is less salient and perceived as less painful than stimulation applied to the forearm, which is in agreement with the larger neuron receptive fields observed for the leg, as well as the overall reduced amount of brain resources destined to process this input. Lastly, the observation that the combined approach (regularized approach) provides the highest prediction gain compared with any other study suggests that indeed the normative population-level predictive maps and idiographic maps provide different information, the combination of which increases predictive accuracy. Overall, the comparison of results across studies suggests that the quality of the data, the neurobiology of the studied phenomenon (more robustly evoked perceptions/sensations/emotions/cognitions will be more reliably predictable than less robust ones, e.g., pain in the forearm vs. in the leg), the characteristics of the population based predictor and the similitude of the to-be-predicted individual to the population from which the normative map was developed, may significantly influence the predictive accuracy obtained with the normative group-level, idiographic and combined (regularized) maps.

With regards to the specific example topic with which we have developed and applied the regularized prediction approach in this study, i.e., pain perception in healthy subjects, a number of potential future directions emerge. One way to increase the power of the GRIP approach over the idiographic approach would be to develop a set of new “population-based” maps that are more specific for the stimulus being applied (e.g., in our case, having different population-based maps for the different stimulation modalities, heat, pressure, chemical, ischemic, incision), and which consider the dynamics of the pain response for different stimulation modalities, durations and frequencies of presentation. For example, though use of the canonical HRF has proven very useful in prediction, it is known that brain responses to both painful heat (Lindquist et al., 2009, 2013; Moulton et al., 2005) and pressure (e.g. Lopez-Sola et al., 2010) have different dynamics from what has been assumed by the canonical hemodynamic response function. On a different note, the population of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

study appears as another important factor to optimize regularized-based prediction models. In our studies, the subjects were all healthy young individuals. It is to be expected that predicting pain perception for subjects that pertain to different populations, such as chronic pain patients, psychiatric patients for which the perception of pain is distorted in various ways, or older subjects, would be much more precise if we derived population-specific normative population-level maps.

As mentioned above, some of our studies included psychological experimental manipulations aimed at modifying the experience of pain potentially via brain circuits that, albeit potentially common across subjects, may not be represented in the normative NPS map. Therefore, the development of a new normative map controlling for the phenomena already accounted for by the NPS (nociceptive, stimulus-intensity dependent brain processing of pain), could account for pain-modulatory brain processes that are mechanistically independent of the more basic nociceptive ones, therefore significantly increasing predictive performance.

Establishing data quality and quality control

Beyond the demonstrated improvement of brain representations by using the GRIP method, this approach can also be used for data quality estimates for a given individual or paradigm. As shown in this set of studies, the predictive quality varies not only between methods (NPS vs. idiographic), but also between studies and between participants. Setting a benchmark for a given paradigm, population or study allows for detecting outliers, which e.g. show a stronger prediction based on the idiographic vs. a group based (e.g. NPS) approach. Equivalently, the GRIP method can be used as a quality control tool within a subject across trials, in this case comparing single trial predictions to e.g., a prediction based on all of the subjects' trials or to a group based prediction. Such a quality control procedure will of course highly depend on the comparison data set(s), but offers the possibility for detecting task-irrelevant variation in a new way.

Single Subject Single Trial Applications

Our results have applications to a number of situations in which brain-based predictions are made from limited amounts of data (e.g., single trials). Brain-based classifiers are increasingly used in brain-computer interface and neuroprosthetic applications, allowing people or animals to manipulate external events via decoding of single-event data. For example, the "Brain Spell" paradigm allows individuals to efficiently write messages without typing or speaking by manipulating their attention, which produces corresponding electroencephalographic activity patterns that are decoded and translated into alphanumeric characters and printed on-screen. Another increasingly popular application is real-time fMRI, in which brief epochs of data (on the order of a few seconds) are prospectively decoded into predictions about mental states. This technique has potential clinical utility; for example, Monti et al., (2010) used an *a priori* pattern based on prior literature on spatial cognition to allow minimally conscious patients to answer yes or no questions, and thereby communicate with the outside world for the first time in years. Patients answered by imagining playing tennis or walking around their house, and classified as yes/no responses based on differences in ventral vs. dorsal posterior cortical activity.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

What is common to all these applications is the need to perform decoding based on very limited amounts of data. Decoding is inherently noisy when based on limited data. In addition, when predictive maps are based on training on a single individual with limited amounts of training data, they are inherently noisy as well, and contribute additional instability to decoding results. In some cases, predictive maps based on limited numbers of trials can be less accurate than using group-level maps; this was the case in a recent study from our group, in which idiographic emotion-predictive maps were less accurate than a predictive map trained across participants (Chang et al. in press). Thus, using population-based maps could help to provide more stable solutions in many cases where it is not possible to obtain large amounts of training data from individual participants.

Acknowledgments

We are grateful for the funding support of NIH, which supported this work under grants R01DA035484 (TDW), 2R01MH076136 (TDW), R01DA027794 (TDW), R01 EB016061 (MAL) and P41 EB015909 (MAL).

Glossary

Predictive map (a.k.a. whole-brain MVPA weight map)
Multivariate brain map, with different predictive weights assigned to each voxel in the brain, optimized to predict a psychological effect or other functional outcome

Idiographic map (a.k.a. individual-subject predictive map)
Predictive map developed on data from a single individual. Predictions are suitable for out-of-sample observations made on the same individual

Population-level map (a.k.a. brain signature)
Predictive map developed on group (multi-subject) data, whose weights constitute estimates of population-level associations between fMRI activity and outcomes. Predictions are suitable for out-of-sample individuals drawn from the same population

References

Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD. Classification of schizophrenia patients based on resting-state functional network connectivity. Front Neurosci. 2013; 7:133. [PubMed: 23966903]

Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005; 26:839–851. [PubMed: 15955494]

Atlas LY, Bolger N, Lindquist MA, Wager TD. Brain mediators of predictive cue effects on perceived pain. The Journal of Neuroscience. 2010; 30:12964–12977. [PubMed: 20881115]

Atlas LY, Lindquist MA, Bolger N, Wager TD. Brain mediators of the effects of noxious heat on pain. PAIN®. 2014; 155:1632–1648. [PubMed: 24845572]

Baucom LB, Wedell DH, Wang J, Blitzer DN, Shinkareva SV. Decoding the neural representation of affective states. Neuroimage. 2012; 59:718–727. [PubMed: 21801839]

Brodersen KH, Wiech K, Lomakina EI, Lin CS, Buhmann JM, Bingel U, Ploner M, Stephan KE, Tracey I. Decoding the perception of pain from fMRI using multivariate pattern analysis. Neuroimage. 2012; 63:1162–1170. [PubMed: 22922369]

Cecchi GA, Huang L, Hashmi JA, Baliki M, Centeno MV, Rish I, Apkarian AV. Predictive Dynamics of Human Pain Perception. Plos Computational Biology. 2012; 8

Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD. A sensitive and specific neural signature for picture-induced negative affect. PLoS Biology. In Press.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Craddock RC, Holtzheimer PE 3rd, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. Magn Reson Med. 2009; 62:1619–1628. [PubMed: 19859933]

Davis T, Poldrack RA. Measuring neural representations with fMRI: practices and pitfalls. Annals of the New York Academy of Sciences. 2013; 1296:108–134. [PubMed: 23738883]

Doehrmann O, Ghosh SS, Polli FE, Reynolds GO, Horn F, Keshavan A, Triantafyllou C, Saygin ZM, Whitfield-Gabrieli S, Hofmann SG, Pollack M, Gabrieli JD. Predicting treatment response in social anxiety disorder from functional magnetic resonance imaging. JAMA Psychiatry. 2013; 70:87–97. [PubMed: 22945462]

Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association. 1975; 70:311–319.

Formisano E, De Martino F, Bonte M, Goebel R. " Who" Is Saying What? Brain-Based Decoding of Human Voice and Speech. Science. 2008; 322:970–973. [PubMed: 18988858]

Freeman J, Brouwer GJ, Heeger DJ, Merriam EP. Orientation decoding depends on maps, not columns. The Journal of Neuroscience. 2011; 31:4792–4804. [PubMed: 21451017]

Friston K, Penny W. Posterior probability maps and SPMs. Neuroimage. 2003; 19:1240–1249. [PubMed: 12880849]

Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J. Classical and Bayesian inference in neuroimaging: theory. Neuroimage. 2002; 16:465–483. [PubMed: 12030832]

Frost MA, Goebel R. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. Neuroimage. 2012; 59:1369–1381. [PubMed: 21875671]

Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, Brammer MJ. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol Psychiatry. 2008; 63:656–662. [PubMed: 17949689]

Gonzalez-Castillo J, Saad ZS, Handwerker DA, Inati SJ, Brenowitz N, Bandettini PA. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proc Natl Acad Sci U S A. 2012; 109:5487–5492. [PubMed: 22431587]

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458:632–635. [PubMed: 19225460]

Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005; 27:83–85.

Haxby JV, Connolly AC, Guntupalli JS. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annual Review of Neuroscience. 2014; 37

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293:2425–2430. [PubMed: 11577229]

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron. 2011; 72:404–416. [PubMed: 22017997]

Heeger DJ, Ress D. What does fMRI tell us about neuronal activity? Nature Reviews Neuroscience. 2002; 3:142–151. [PubMed: 11836522]

Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y. Neural decoding of visual imagery during sleep. Science. 2013; 340:639–642. [PubMed: 23558170]

Huth AG, Nishimoto S, Vu AT, Gallant JL. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron. 2012; 76:1210–1224. [PubMed: 23259955]

IOM. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research. Institute of Medicine; Washington, DC: 2011.

James, W.; Stein, C. Estimation with quadratic loss. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability; 1961. p. 361-379.

Johnson JD, McDuff SG, Rugg MD, Norman KA. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron. 2009; 63:697–708. [PubMed: 19755111]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nat Neurosci. 2005; 8:679–685. [PubMed: 15852014]

Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA. Identifying Emotions on the Basis of Neural Activation. Plos One. 2013; 8:e66032. [PubMed: 23840392]

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452:352–355. [PubMed: 18322462]

Kuhl BA, Rissman J, Chun MM, Wagner AD. Fidelity of neural reactivation reveals competition between memories. Proc Natl Acad Sci U S A. 2011; 108:5903–5908. [PubMed: 21436044]

Lindquist MA, Gelman A. Correlations and multiple comparisons in functional imaging: a statistical perspective (Commentary on Vul et al., 2009). Perspectives on Psychological Science. 2009; 4:310–313. [PubMed: 26158969]

Lindquist MA, Loh JM, Atlas LY, Wager TD. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. Neuroimage. 2009; 45:S187–S198. [PubMed: 19084070]

Lopez-Sola M, Pujol J, Hernandez-Ribas R, Harrison BJ, Ortiz H, Soriano-Mas C, Deus J, Menchon JM, Vallejo J, Cardoner N. Dynamic assessment of the right lateral frontal cortex response to painful stimulation. Neuroimage. 2010; 50:1177–1187. [PubMed: 20080188]

Marquand A, Howard M, Brammer M, Chu C, Coen S, Mourao-Miranda J. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage. 2010; 49:2178–2189. [PubMed: 19879364]

Marquand AF, Brammer M, Williams SC, Doyle OM. Bayesian multi-task learning for decoding multi-subject neuroimaging data. Neuroimage. 2014; 92:298–311. [PubMed: 24531053]

Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA. Improving Reliability of Subject-Level Resting-State fMRI Parcellation with Shrinkage Estimators. 2014 arXiv preprint arXiv:1409.5450.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. Predicting human brain activity associated with the meanings of nouns. Science. 2008; 320:1191–1195. [PubMed: 18511683]

Monti MM, Vanhaudenhuyse A, Coleman MR, Boly M, Pickard JD, Tshibanda L, Owen AM, Laureys S. Willful modulation of brain activity in disorders of consciousness. New England Journal of Medicine. 2010; 362:579–589. [PubMed: 20130250]

Moulton EA, Keaser ML, Gullapalli RP, Greenspan JD. Regional intensive and temporal patterns of functional MRI activation distinguishing noxious and innocuous contact heat. Journal of neurophysiology. 2005; 93:2183–2193. [PubMed: 15601733]

Mumford JA, Turner BO, Ashby FG, Poldrack RA. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage. 2012; 59:2636–2643. [PubMed: 21924359]

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology. 2011; 21:1641–1646. [PubMed: 21945275]

Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in cognitive sciences. 2006; 10:424–430. [PubMed: 16899397]

Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental States across individuals. Psychological Science. 2009; 20:1364–1372. [PubMed: 19883493]

Racine J. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of econometrics. 2000; 99:39–61.

Rissman J, Gazzaley A, D'Esposito M. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage. 2004; 23:752–763. [PubMed: 15488425]

Rissman J, Greely HT, Wagner AD. Detecting individual memories through the neural decoding of memory states and past experience. Proc Natl Acad Sci U S A. 2010; 107:9849–9854. [PubMed: 20457911]

Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA. Using FMRI brain activation to identify cognitive states associated with perception of tools and dwellings. Plos One. 2008; 3:e1394. [PubMed: 18167553]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Shou H, Eloyan A, Nebel MB, Mejia A, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA, Crainiceanu CM. Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI. Neuroimage. 2014; 102:938–944. [PubMed: 24879924]

Siegle GJ, Carter CS, Thase ME. Use of FMRI to predict recovery from unipolar depression with cognitive behavior therapy. Am J Psychiatry. 2006; 163:735–738. [PubMed: 16585452]

Su SC, Caffo B, Garrett-Mayer E, Bassett SS. Modified test statistics by inter-voxel variance shrinkage with an application to fMRI. Biostatistics. 2009; 10:219–227. [PubMed: 18723853]

Tracey I. Can neuroimaging studies identify pain endophenotypes in humans? Nature reviews. Neurology. 2011; 7:173–181. [PubMed: 21304481]

Wager TD, Atlas LY, Leotti LA, Rilling JK. Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. The Journal of Neuroscience. 2011; 31:439–452. [PubMed: 21228154]

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

Wager TD, Nichols TE. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. Neuroimage. 2003; 18:293–309. [PubMed: 12595184]

Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, Barker GJ, Bokde AL, Büchel C, Carvalho FM. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature. 2014; 512:185–189. [PubMed: 25043041]

Woo CW, Koban L, Kross E, Lindquist MA, Banich MT, Ruzic L, Andrews-Hanna JR, Wager TD. Separate neural representations for physical pain and social rejection. Nat Commun. 2014; 5:5380. [PubMed: 25400102]

Woo CW, Roy M, Buhle JT, Wager TD. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. PLoS Biol. 2015; 13:e1002036. [PubMed: 25562688]

Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA. Greater neural pattern similarity across repetitions is associated with better memory. Science. 2010; 330:97–101. [PubMed: 20829453]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

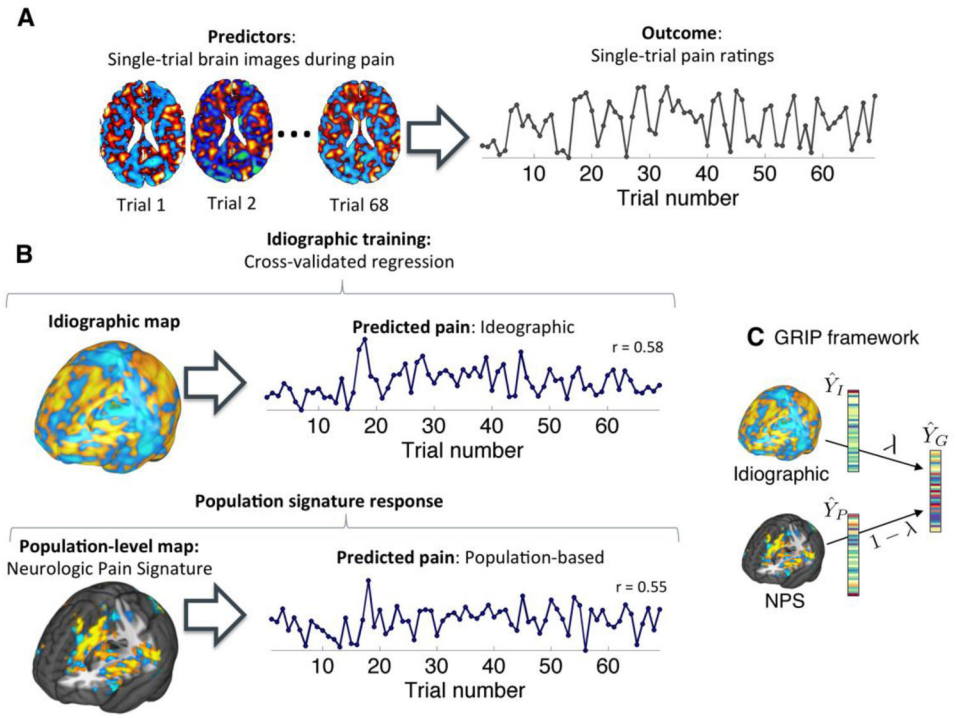

Figure 1.

Schematic representation of the group-regularized individual prediction (GRIP) framework as applied to pain data. This illustrates the framework and procedure, which can in principle be applied to data from any domain or task. **(A)** Single-trial images are estimated using a first-level general linear model with a separate regressor per trial. These images, which are composed of pain-related activation estimates from each gray matter voxel, are used to predict trial-by-trial reports of pain intensity. **(B)** Idiographic maps for each individual participant are constructed by regressing pain reports on single-trial voxel values. We use LASSO-PCR (Wager et al., 2013; see text) with leave-one-run-out cross-validation, so that all predictions are out-of-run. Population-based predictions are obtained by applying the Neurologic Pain Signature (NPS; Wager et al., 2013), a normative signature trained on independent data (from Study 1) to predict pain in out-of-sample individuals. In both idiographic and population-based methods, predicted responses can be obtained by calculating the dot product between the predictive map and an input image; this is the standard way of calculating predicted responses in linear regression. In the case of idiographic prediction, the predicted pain intensities ($\hat{Y}_I$) are based on weights obtained from training data from other runs. In the case of population-based prediction, the predicted pain intensities ($\hat{Y}_P$) are based on weights from the NPS. Weight maps can be applied to any image (a contrast map, a single-trial image, or single time-point image), though here they are applied to single-trial images. **(C)** To maximize prediction accuracy, GRIP predictions ($\hat{Y}_G$) are generated by combining trial-by-trial predictions, based on subject-specific idiographic predictive maps ($\hat{Y}_I$) with predictions based on population-level predictive maps ($\hat{Y}_P$). The relative weight allotted to each map is controlled by the shrinkage factor λ.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

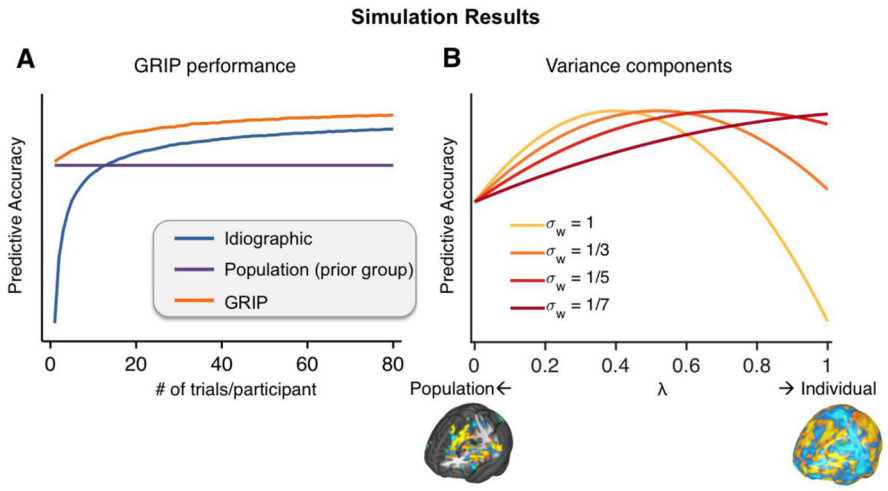

Figure 2.

Simulation results for group-regularized individual prediction (GRIP) based on Empirical Bayes regularization. **(A)** The results of a theoretical simulation showing single-trial prediction accuracy as a function of the number of trials per participant. Here we assume a fixed variance magnitude within and between participants based on reasonable values from prior studies. Group predictions use a static, *a priori* map and are thus not affected by the number of trials, while idiographic predictions are very poor with low amounts of data but improve with more training trials per participant. As a result, the GRIP prediction, which is a combination of both the group and idiographic maps, will also increase with more training trials per participant and eventually will converge with the idiographic prediction. **(B)** The results of a theoretical simulation showing single trial prediction accuracy as a function of the shrinkage factor λ, which varies from 0 (100% group estimate) to 1 (100% individual estimate). When the within participant variance is high, predictive accuracy will peak for λ values closer to 0, whereas when the within participant variance decreases, prediction accuracy peaks for λ values closer to 0.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

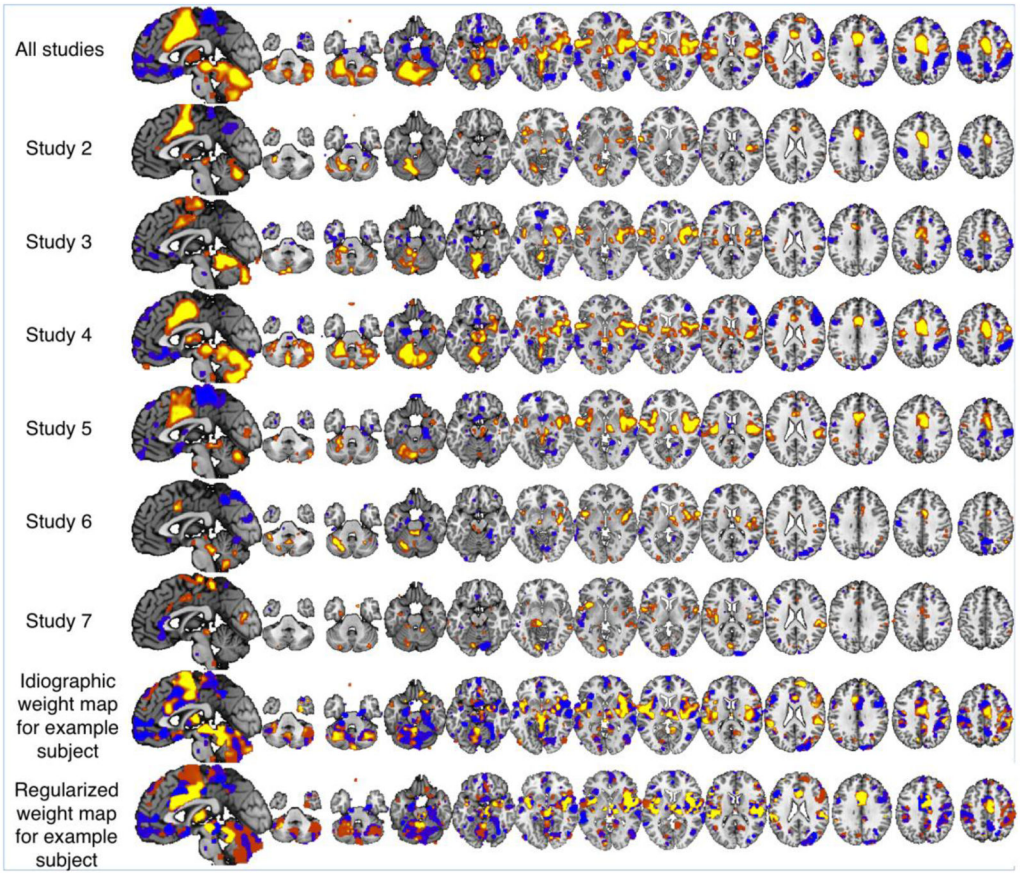

Figure 3.

Idiographic predictive maps (weights) for each study and results for an example subject. Shown are the average brain weights across all subjects and studies in the test datasets (Studies 2–7, N = 180, top row), as well as study-specific averages for each study (bottom rows). Study 1 was the training dataset for the population-level map used to regularize the other studies (the Neurologic Pain Signature; (Wager et al., 2013), so is not included here. In addition, an example subject (from Study 2, with idiographic predictive accuracy r = .5) is included both before and after shrinkage to the study mean. Yellow/orange values indicate positive weights; greater activation predicts increased reported pain. Blue values indicate negative weights; greater activation predicts less reported pain. “All studies” refers to the average weight across studies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

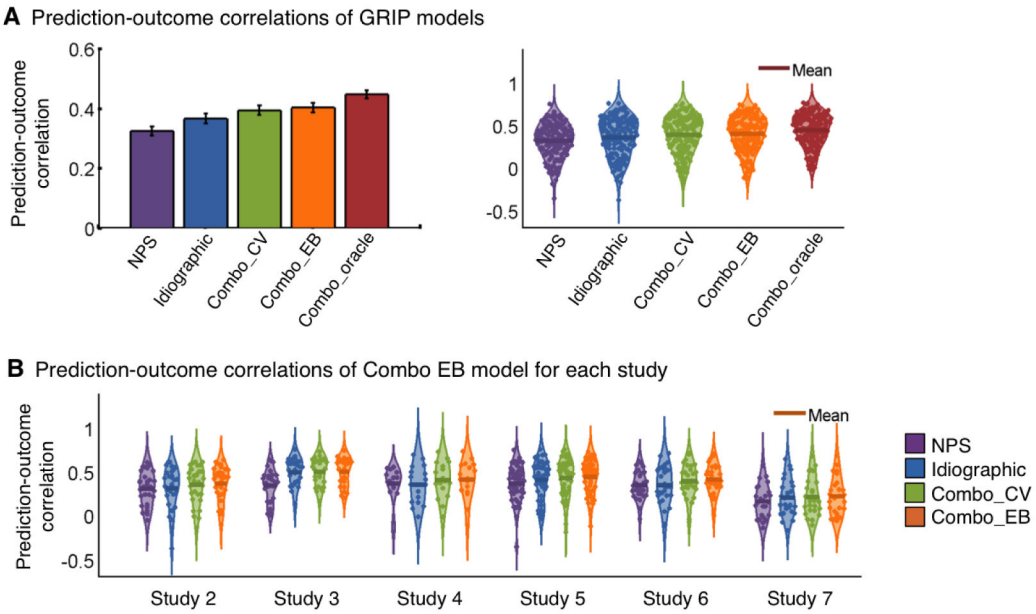

Figure 4.

Predictive accuracy as a function of regularization method and study using violin plots (i.e., a box plot with a rotated kernel density plot appearing on each side). **(A)** Prediction accuracies (n = 180) as measured by the correlation between pain ratings (outcome) and predictions computed using the Neurologic Pain Signature (NPS) population-level map, an idiographic map, and three regularized combination maps (using the cross-validated (CV), empirical Bayes (EB) and oracle approaches). The oracle method provides an upper bound on linear regularization. The EB regularization is better than any other model besides the oracle, and so it the preferred method. **(B)** Prediction-outcome correlations for the NPS, idiographic and EB maps for each of the six studies included. All methods predicted single-trial pain reports substantially above chance (0) in all studies. Using idiographic predictive maps outperformed the NPS for 3 of the 6 studies, and predictions were approximately equally accurate for the other 3 studies. The EB method was the best method for all 6 studies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

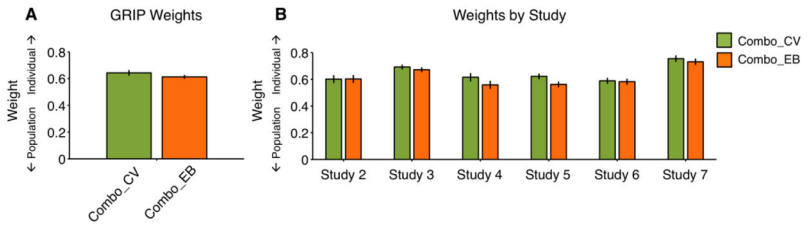

Figure 5.

Regularization parameter estimates. **(A)** Average shrinkage factors (λ, n = 180) for the combined empirical Bayes (EB) and cross-validated (CV) approaches. **(B)** Average shrinkage factors for the empirical Bayes (EB) and cross-validated (CV) approaches for each of the 6 studies. The results indicate a relatively uniform shrinkage value across studies, with a balance between population and individualized (idiographic) predictive maps and a moderate bias towards the individualized maps. Notably, shrinkage factors in Study 7 favor individual maps to a greater degree; this study was the only one to deliver a different type of painful heat on a different body site (leg) from the population map. Error bars correspond to ± 1 s.d.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

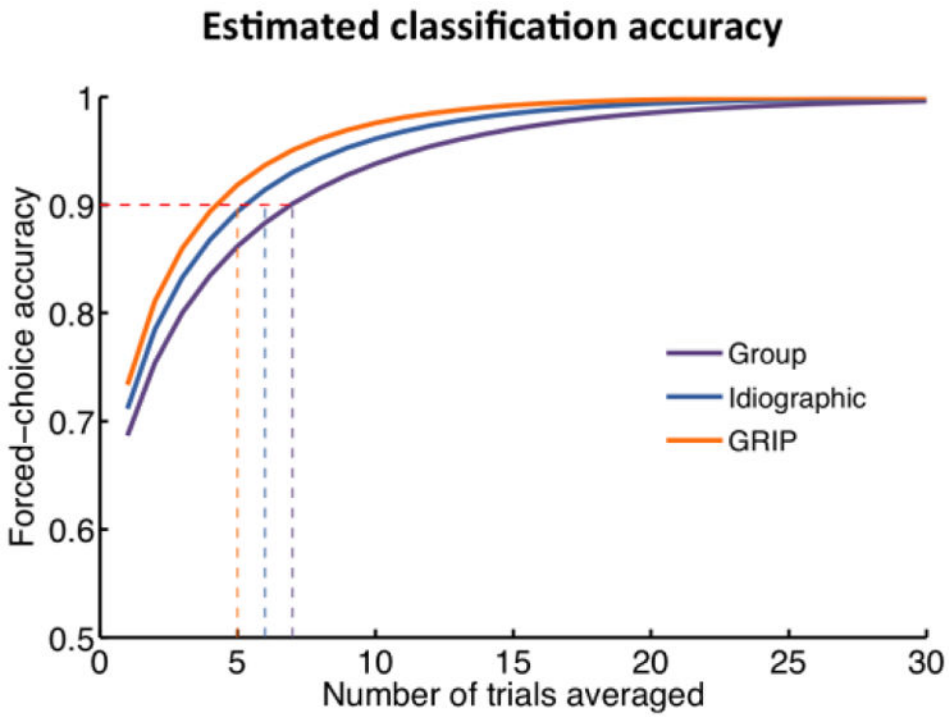

Figure 6.

Prediction accuracy for high vs. low pain forced-choice classification for NPS, idiographic, and combined empirical Bayes (EB) approaches as a function of the number of trials averaged. Accuracy depends on the effect size, or how strongly the predictive map is related single-trial outcome measures, and the number of trials averaged. Forced-choice accuracy is the proportion of correct decisions made when two maps from a single participant are compared and they differ by one standard deviation in outcome (pain report).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

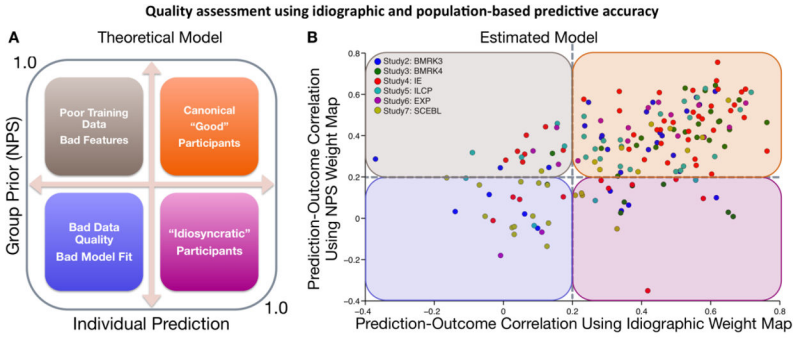

Figure 7.

Relationship between the prediction accuracies obtained using idiographic and population-level maps. Prediction accuracies are measured by the trial-by-trial correlation between pain ratings (outcome) and brain-based predictions. Every dot corresponds to one participant. Each study is coded with a different color. A threshold of r = 0.2 was used to define four quadrants of prediction accuracy (quadrant 1 (top-right): high-high; quadrant II: high-low, quadrant III: low-low, quadrant IV: low-high). Though a heuristic, it is based on the fact that on average, within-person correlations of 0.2 will be significant in our studies.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Demographics

Study◆	Sample Size	Sex	Mean age in Years (Std. Deviation)	Prior publications
NPS Training Data				
Study 1 (NSF)	26	9 F	27.8	Atlas et al. (2014), Pain; Wager et al. (2013) NEJM
NPS Testing Data				
Study 2 (BMRK3)	33	22 F	27.9 (9.0)	Woo et al. (2015), PLOS Biology Wager et al. (2013) NEJM
Study 3 (BMRK4)	28	10 F	25.2 (7.4)	Krishnan et al. (*Under Review*)
Study 4 (IE)	50	27 F	25.1 (6.9)	Roy et al. (2014), Nature Neuroscience
Study 5 (ILCP)	29	16 F*	20.4 (3.3)**	Schmidt et al. (*In Prep.*)
Study 6 (EXP)	17	9 F	25.5	Atlas et al. (2010), Journal of Neuroscience
Study 7 (SCEBL)	26	11 F	28 (9.3)	Koban et at. (*In Prep.*)

Note.

◆Internal study codes to facilitate tracking of datasets;

*Gender of one participant is unknown;

**Age of one participant is unknown.

Publications include: (Atlas et al., 2010a; Atlas et al., 2014a; Roy et al., 2014; Wager et al., 2013; Woo et al., 2015b).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Stimulation Parameters

Study	Intensities	Mean Temperature by Intensity Level (Within Subject SE)	Rating scale	Mean Ratings by Intensity Level (Within Subject SEM)
NPS Training Data				
Study 1 (NSF)	N, L, M, H (Calibrated)	40.8, 43.1, 45.1, 47.0 (0.16)	0–8 VAS (0, no sensation; 1, non-painful warmth; 2, low pain; 5, moderate pain; 8, maximum tolerable pain)	2.0, 2.8, 4.2, 6.6 (0.14)
NPS Testing Data				
Study 2 (BMRK3)	6 levels (Fixed)	44.3, 45.3, 46.3, 47.3, 48.3, 49.3	0–100 VAS	49.1, 56.6, 74.3, 99.4, 133.0, 159.3 (3.12)
Study 3 (BMRK4)	L, M, H (Fixed)	46.0, 47.0, 48.0	0–100 VAS (0, no sensation; 1.4, barely detectable; 6.1, weak; 17.2, moderate; 35.4, strong; 53.3, very strong; 100, strongest imaginable sensation)	UL: 31.7, 40.5, 53.6 (0.9787) LL: 31.5, 40.2, 53.3 (0.96)
Study 4 (IE)	L, M, H (Fixed)	46.0, 47.0, 48.0	0–100 VAS (0, no pain; 100, worst imaginable pain)	29.4, 38.9, 51.9 (0.64)
Study 5 (ILCP)	L, H (Calibrated)	44.7, 46.7 (0)	0–8 VAS (no pain to worst pain imaginable)	24.3, 46.7 (1.14)
Study 6 (EXP)	L, M, H (Calibrated)	41.2, 44.4, 47.2 (0.21)	0–8 VAS (0, no sensation; 1, non-painful warmth; 2, low pain; 5, moderate pain; 8, maximum tolerable pain)	2.5, 4.3, 7.4 (0.13)
Study 7 (SCEBL)	L, M, H (Fixed)	48, 49, 50	0–100 VAS (0, no pain; 100, worst imaginable pain)	26.0, 33.3, 40.4 (1.12)

Note: Heat/pain levels: N = Nonpainful, L = Low, M = Medium, H = High. Sites of stimulation: UL = Upper Limb, LL = Lower Limb. VAS = visual analogue scale.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Task Characteristics

Study	Duration (seconds)	Inter-heat interval (seconds)	Locations (number of sites)	Range of Number of Trials Per Subject	Mean proportion of trials excluded (Std. Deviation)	Other experimental manipulations
NPS Training Data						
Study 1 (NSF)	10	38	Arm (3)	35–48	0.08 (0.07)	Masked emotional faces evenly crossed with temperature
NPS Testing Data						
Study 2 (BMRK3)	12.5	20.5–28.5	Arm (2)	97	0.1 (0.04)	Cognitive self-regulation up and down
Study 3 (BMRK4)	11	25–27	Arm (4), Foot (4)	81	0.08 (0.06)	Heat-predictive visual cues (low, medium, or high)
Study 4 (IE)	11	36–38	Arm (6)	48	N/A	Heat-predictive visual cues; placebo manipulation
Study 5 (ILCP)	10	17–25	Arm (2)	64	0.05 (0.03)	Agency (make choice, observe choice), Certainty (80% low pain, 50% low pain)
Study 6 (EXP)	10	38	Arm (4)	61–64	0.03 (0.04)	Heat-predictive auditory cues
Study 7 (SCEBL)	1.85	26–37	Leg (6)	96	0.04 (0.03)	Heat-predictive visual cues (low or high) and unreinforced social information

Note: The exclusion criterion was a high variance inflation factor.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Acquisition Parameters

Study	Study Location	Scanner Details	EPI Parameters	Voxel Size (mm^3)	Acquisition Parameters	Discarded Volumes	Stimulus Software	Analysis Software
NPS Training Data								
Study 1 (NSF)	Columbia	1.5T GE Signa TwinSpeed Excite HD	TR = 2000 ms TE = 34 ms FOV = 224 mm Matrix = 64×64	3.5×3.5× 4.0	24 slices	5	E-Prime	SPM 8
NPS Testing Data								
Study 2 (BMR K3)	Columbia	3T Phillips Achieva TX	TR = 2000 ms TE = 20 ms FOV = 224 mm Matrix = 64×64	3.0×3.0× 3.0	42 Slices Interleaved SENSE = 1.5	4	E-Prime	SPM 8
Study 3 (BMR K4)	CU Boulder	3T Siemens Tim Trio	TR = 1300 ms TE = 25 ms FOV = 220 mm Matrix = 64×64 Flip Angle = 50°	3.4×3.4× 3.4	26 Slices Interleaved iPAT = 2	6	Matlab	SPM 8
Study 4 (IE)	CU Boulder	3T Siemens Tim Trio	TR = 1300 ms TE = 25 ms FOV = 220 mm Matrix = 64×64 Flip Angle = 75°	3.4×3.4× 3	26 Slices Interleaved iPAT = 2	6	E-Prime	SPM 8
Study 5 (ILCP)	CU Boulder	3T Siemens Tim Trio	TR = 1980 ms TE = 25 ms FOV = 220 mm Matrix = 64×64 Flip Angle = 75°	3.4×3.4× 3	35 Slices Interleaved iPAT = 0	5	Matlab	SPM 8
Study 6 (EXP)	Columbia	1.5T GE Signa TwinSpeed Excite HD	TR = 2000 ms TE = 40 ms FOV = 224 mm Matrix = 64×64 Flip Angle = 84°	3.5×3.5× 4.55	24 Slices	5	E-Prime	SPM 5
Study 7 (SCE BL)	CU Boulder	3T Siemens Tim Trio	TR = 1300 ms TE = 25 ms FOV = 220 mm Matrix = 64×64 Flip Angle = 50°	3.4×3.4× 3.4	26 Slices Interleaved iPAT = 2	3	E-Prime	SPM 8

Note. TR = Time to Repeat; TE = Time to Echo; FOV = Field of View