Learning to identify CNS drug action and efficacy using multi-study fMRI data

Eugene. P. Duff[1*], William Vennart[2], Richard Wise[3], Matthew A. Howard[4], Richard E. Harris[5], Michael Lee[1], Karolina Wartolowska[1], Vishvarani Wanigasekera[1], Frederick J. Wilson[2], Mark Whitlock[2], Irene Tracey[1], Mark W. Woolrich[1,6], Stephen M. Smith[1]

[1]FMRIB Centre, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX39DU, United Kingdom

[2]Pfizer Ltd., Cambridge CB21 6GS, United Kingdom

[3]Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, United Kingdom

[4]Department of Neuroimaging, Institute of Psychiatry, King's College, London SE58AF, United Kingdom

[5]Department of Anesthesiology, University of Michigan, Ann Arbor, MI 48105, USA.

[6]Oxford University Centre for Human Brain Activity (OHBA), Department of Psychiatry, University of Oxford, Oxford, United Kingdom

*To whom correspondence should be addressed, email: eugene.duff@ndcn.ox.ac.uk

One Sentence Summary: Existing functional brain imaging datasets were used to identify neural signatures that confirm pharmacological action and predict clinical efficacy of test compounds.

Abstract:

The therapeutic effects of centrally acting pharmaceuticals can manifest gradually and unreliably in patients, making the drug discovery process slow and expensive. Biological markers providing early evidence for clinical efficacy could help prioritize development of the more promising drug candidates. A potential source of such markers is functional magnetic resonance imaging (fMRI), a noninvasive imaging technique that can complement molecular imaging. fMRI has been used to characterize how drugs cause changes in brain activity. However, variation in study protocols and analysis techniques has made it difficult to identify consistent associations between subtle modulations of brain activity and clinical efficacy. In this work we present and validate a general protocol for functional imaging–based assessment of drug activity in the central nervous system. The protocol uses machine-learning methods and data from multiple published studies to identify reliable associations between drug-related activity modulations and drug efficacy, which can then be used to assess new data. A proof-of-concept version of this approach was developed and is shown here for analgesics (pain medication), and validated with 8 separate studies of analgesic compounds. Our results show that the systematic integration of multi-study data permits the generalized inferences required for drug discovery. Multi-study integrative strategies of this type could help optimize the drug discovery and validation pipeline.

Introduction

Central nervous system (CNS) drug failure rates are high throughout the drug-development cycle, with initial human trials a common point of failure *(1, 2)*. Assessment of the therapeutic potential of CNS drug candidates in humans can be difficult and expensive, with efficacy unreliable, hard to measure, and slow to manifest. Ideally, ultimately unsuccessful drugs would fail earlier in the development process (before moving into patient trials). Biomarkers that prioritize candidate compounds prior to large-scale clinical trials have the potential to substantially improve the productivity and cost-effectiveness of drug development *(1–3)*.

Functional magnetic resonance imaging (fMRI) can image neural responses and their pharmacological modulation through the blood oxygen level–dependent (BOLD) contrast mechanism. fMRI is capable of characterizing the effects on the CNS of drugs associated with conditions such as chronic pain *(3)*, schizophrenia *(4)*, and obesity *(5)*. In the area of pain, fMRI studies have identified specific reductions in pain responses with the analgesics remifentanil *(6)*, alfentanil *(7)*, and ketamine *(8)*; gabapentin-induced changes in deactivations *(9)*; and reductions

in resting-state functional connectivity with opiates *(10, 11)*. Neural correlates of many other clinically relevant aspects of CNS disease have been identified including clinical scores *(12, 13)*, anxiety and fear *(14)*, the placebo response *(15)*, and sensitization *(9)*. Importantly, fMRI has shown potential for distinguishing effective from non-effective compounds *(16, 17)*, and for predicting the clinical effects of drugs *(15, 18, 19)*. These capabilities suggest that fMRI could provide a complementary, non-invasive adjunct to molecular imaging for drug discovery, detecting drug-related modulations of brain activity (referred to here as "pharmacodynamic action") predictive of efficacious drug action (*4*, *5*).

Despite the promise of fMRI, progress towards standard, validated procedures for assessing the potential efficacy of new CNS compounds has been slow *(17)*. In part, this is due to the difficulty of using fMRI to make inferences regarding the likelihood that a drug will be successful in a specific clinical domain ("clinical efficacy"). fMRI does not quantify physiological variables directly associated with drug action, so identifying evidence for the efficacy of compounds must be based upon empirically established associations between brain activity patterns and measurable clinical variables such as treatment outcome. It is important that imaging methods are able to provide predictive capabilities beyond what can be obtained from clinical measures alone. Direct brain correlates of available behavioral and clinical measurements, which may be affected by factors unrelated to long-term efficacy, will not necessarily provide substantial additional predictive value for drug assessment. One approach that may provide complementary data is to identify common effects of existing compounds or treatments that have already been shown to be efficacious in clinical trials; this approach could identify brain activity modulations suggestive of efficacy that occur earlier or more reliably than changes in external clinical measures.

RESULTS

Study design and rationale

We designed a procedure for performing imaging-based drug assessment, intended to detect drug-related modulation of brain activity (pharmacodynamic effects) and evidence that this modulation suggests clinical efficacy. We focused on defining a protocol for placebo-controlled crossover designs. These low-cost studies, typically using 5-40 subjects, are widely used in pharmacological imaging, and have been proposed as a screening stage for CNS compounds prior to phase IIa clinical trials (1). We assumed that many aspects of experimental design would vary across studies, with study designs being optimized for specific clinical conditions and experimental questions. Our approach uses image-based mega-analysis *(20)* and multivariate pattern analysis (MVPA) methods (*14*, *20*) to identify evidence for data and stimulus validity (i.e. quality assurance), pharmacodynamic modulation of brain activity, and evidence that these modulations suggest efficacy. Detection of pharmacological effects and evidence for clinical efficacy was achieved by testing different algorithms trained to discriminate the effects of drug from placebo.

We tested a proof-of-concept implementation of the protocol on a set of 8 clinical fMRI datasets and 6 control datasets derived from the same studies [table 1, *(6, 9, 18, 21–25)*]. These studies investigated the effects of analgesic compounds on brain responses to painful stimuli. The assessment procedure was applied to each study individually, using data from the remaining studies to identify signatures of responses to pain and drug modulations.

General fMRI drug assessment procedure and application to analgesic studies

Our fMRI drug assessment protocol has three assessment stages, each addressing distinct questions: Quality assurance (Fig. 1A)—are the data and model of sufficient quality to identify the anticipated effects? Pharmacodynamic effect (Fig. 1B)—is the compound modulating brain activity? Evidence for clinical efficacy (Fig. 1C; fig. S1)—is there evidence for modulation of brain activity that has been shown to be associated with clinical efficacy? These assessment stages could be applied in a sequential manner using predefined decision rules.

Quality assurance. Quality assurance (QA) is necessary to ensure that the acquired fMRI data and analysis strategies provide good prospects for identifying drug effects, should they exist. QA should include a range of image quality and registration assessments as well as assessments of model validity and statistical procedures. To assess drug-related modulations of responses, it is important that the modeled responses accurately reflect the processes of interest, such as pain. In the proof-of-concept assessments of analgesics, we tested a QA assessment procedure that determined whether the pain responses recorded in the test study were similar to those in a set of existing studies (Fig. 1A). A set of stimulus-response parameter maps from each study were generated using a general linear model (GLM) that included standardized regressors for the painful stimuli. A further GLM was used to model differences between responses in the test study compared to other studies. The QA assessment flagged a study as potentially problematic if it showed lower pain-related responses (across the combined drug and placebo sessions) compared to other studies in regions that consistently showed responses to painful stimuli (Supplementary Methods).

Assessment of pharmacodynamic effects. Identifying evidence that a drug is modulating brain responses ("pharmacodynamic effect") can increase confidence in a compound when moving forward to large-scale clinical trials *(1, 2)*. Assessment of pharmacodynamic

effects is particularly important where target engagement is questioned, dosing is uncertain, or the mode of action is poorly understood. Machine-learning classification methods can provide sensitive detection of subtle and spatially extended patterns of pharmacodynamic effects on brain responses. The presence of a drug modulation can be assessed using cross-validation to quantify the reliability with which a classifier can distinguish drug from placebo sessions in held-out data (e.g., subjects not used in the training of the classifier) (Fig. 1B). The resulting prediction accuracies and associated p-values provide a simple measure for assessing the presence of drug effects.

In our proof-of-concept application, a forced-choice support vector machine (SVM) was used for classification, using features derived from an Independent Component Analysis (ICA) based decomposition of the task response maps *(26–29)*. This classifier was trained to predict which of the sessions was associated with the test analgesic. A leave-one-subject-out cross-validation scheme generated predictions for all subjects. Reliable identification of the drug session indicated evidence for a pharmacodynamic effect (Supplementary methods).

Evidence for clinical efficacy. fMRI studies of the target disease, existing effective compounds, or brain systems targeted by the candidate compound, can provide indirect "proxy" evidence linking specific patterns of response modulations to increased likelihood of clinical efficacy for the test compound. The third stage of the drug assessment procedure consists of quantitative tests for such modulations.

For our primary assessment for evidence of clinical efficacy with the proof-of-concept protocol for analgesic studies, we used an approach that aimed to identify a common signature of the effects of established compounds (Fig. 1C). Although different efficacious compounds will

have varying modes of action, it is likely that many have commonalities in their action, or will produce similar modulations of downstream neural activity.

A forced-choice SVM algorithm was trained to distinguish brain responses to pain in the presence of analgesics from responses during placebo, using data from a set of studies of established analgesics (Table 1, a to h). This classifier was tested on the target compound, to determine whether the effects of the target compound had similar enough action to the signature of analgesic action to be discriminated from placebo. This procedure was applied to parameter maps generated from individual subjects, producing prediction accuracies that could be compared to the results of the pharmacodynamic effect assessment.

Preexisting efficacious compounds will not always be available for analysis. Evidence of efficacious drug action might also be obtained by training a classifier to identify modulations of brain responses associated with improvements in disease symptoms; for example, by training it to distinguish lower from higher pain levels *(19)*. A second clinical efficacy assessment approach was tested where responses to different intensities of painful stimuli were used to train a classifier. This classifier was then tested for its ability to identify the drug condition as the lower pain state when compared to placebo (Supplementary methods, fig. S1).

Datasets for proof-of-concept validation

The proof-of-concept protocol was tested on eight placebo-controlled crossover fMRI studies of the effects of established analgesics on responses to pain (Table 1, a to h)*(6, 9, 18, 21–25)*. All of the studies investigated how brain responses to short-lasting painful stimuli, such as brief laser pulses, were modulated by analgesics, in patients or healthy controls, with separate scanning

sessions for the placebo and active drug conditions. The stimuli were always painful, but varied in their timing, location, and quality. Several studies involved multiple types or levels of painful stimuli (e.g. thermal, punctate, brush, squeeze) (table 1, b-e and g-h). The studies investigated analgesics from a variety of compound classes, and varied in the MRI scanners, disease states, dosing regimens, and numbers of subjects used (Table 1). Being capable of using heterogeneous datasets ensures that a high proportion of existing studies can be used to inform and validate assessments, and that assays can be developed in the context of evolving acquisition and experimental protocols.

The protocol was tested for its robustness to false-positives using six datasets measuring effects not associated with analgesia (Table 1, i to n). Three of these control studies were generated from the test studies by replacing the active drug-session data with data from a second placebo or baseline session that had also been obtained in the study. Also assessed were control data where the painful stimuli were replaced with stimuli activating systems unrelated to analgesia, such as visual stimuli. We expected these data to be flagged in the QA assessment, as the stimulus responses were unlikely to resemble responses to painful stimuli.

When the assessment protocol was applied to a study, the remaining analgesic studies were used as training data for QA and clinical efficacy assessments. Training sets excluded any studies assessing the same compound as the test study (so inferences would be valid for unstudied compounds), and studies that showed no evidence for pharmacodynamic effects.

Decision rules for proof-of-concept assessments

We defined some provisional decision rules to structure our proof-of-concept assessments. The choice of decision rules depends on the specific drug development context, taking into account existing evidence for the test compound, the promise of alternative compounds, and confidence in the imaging protocol.

For QA, a test study was flagged as potentially problematic if the average pattern of pain responses deviated from those seen in other studies; specifically, if it showed significantly lower responses to pain compared to those seen in the set of existing studies, in areas showing consistent responses in the other studies. If a study was flagged, we assessed data and modeling to determine the source of the discrepancy. For the drug effect and efficacy assessments, compounds were flagged as showing positive evidence for a drug effect if the whole-brain response maps measured in drug sessions could be discriminated from placebo using a classifier (Supplementary methods) at a rate greater than P=0.05 chance level (one-sided binomial test). As even modest evidence for a drug effect can provide substantial value to decision-making, discrimination at P-values ranging from 0.05 to 0.20 were given a qualified positive flag. We gave a study an overall "Go" outcome if it passed all of the decision stages, and a "Qualified Go" outcome if there was some evidence of pharmacodynamic effects or clinically efficacious action (Fig. 1C).

Anomalous stimulus responses identified by multi-study QA assessment. One study, of pregabalin (Table 2, b)*(24)*, initially failed the QA component of the proof of concept assessment, showing no positive responses to pain (fig. S1C). Average time courses in key pain brain regions were generated using anatomical ROIs (see "Validation of prediction methods") and average responses from all subjects and sessions were extracted (n = 23, fig. S1B). Investigation of time-courses of the fMRI BOLD pain-related responses determined that BOLD

pain responses were much briefer than modeled in the GLM (fig. S2, A and B). The data were re-analyzed using a GLM using an additional component modeling a transient response to the onset of the painful stimulus, which corrected this modeling error (fig. S2, C and D). Diagnosing and updating the modeling in this way did not introduce biases into the drug assessment, as the average pain responses across conditions were orthogonal to the drug effect.

Proof of concept fMRI protocol consistently identified pharmacodynamic effects of analgesics

The results of applying the proof-of-concept procedure to the fMRI test datasets (Table 1) are shown in Table 2. Moderate to strong evidence for an analgesic drug like effect was identified for all analgesic compounds, while all control assessments of placebo conditions identified no evidence.

All analgesic studies consistently identified a pain response. Regions that consistently responded to the painful stimuli included insula, cingulate, and sensory cortices (Fig. 2A). The control assessments involving innocuous auditory or visual stimuli were flagged as producing responses that did not correspond to pain responses.

All but one of the tested analgesic fMRI studies showed evidence for a pharmacodynamic effect. In the successful drug assessments, analgesics were distinguished from placebo at rates ranging from 70 to 92%. When these studies were analyzed in individual GLM mapping analyses (with responses to pain in drug sessions subtracted from responses in placebo sessions), reduced brain responses during the drug sessions were evident in many studies, but the locations

of significant clusters were not consistent, indicating that standard spatial localization analyses did not have the sensitivity to identify consistent effects (fig. S3). The study assessing the effects of tramadol on the brain responses to pain of post-traumatic neuropathic pain (PTNP) patients (Table 2, d), showed no evidence of a pharmacodynamic effect in the classification and GLM analysis. No pharmacodynamic effects were identified in the control studies where a second placebo replaced the drug condition, [Table 2, (i-j)] *(6, 18)*. Remifentanil and naproxen modulated responses to visual and auditory stimuli [Table 2, (l-n)]*(21, 22)*.

Identification of evidence of clinically efficacious (analgesic) effects

The assessment for clinical efficacy found moderate to strong evidence for an analgesic-like effect on responses for all but one of the analgesic studies [Table 2, (a-h)]. Successful identification of analgesic sessions ranged from 57% to 83%, tending to be equal or less than study-specific pharmacodynamic prediction accuracies, except for tramadol. Despite modulating pain ratings and responses, THC was not reliably discriminated from placebo by the clinical efficacy assessment, with accuracy of only 57% (ci: 36-76%). THC has a dissociative effect on pain perception, so may not produce the low-level modulations of pain responses produced by other analgesics. When sessions involving analgesic compounds were replaced by placebo or baseline sessions, prediction accuracies for the efficacy assessment algorithm were at chance levels [Table 2, (l and m)]. The algorithm also did not consistently identify the drug modulation of responses to non-painful visual and auditory stimuli as signs of potential efficacy [Table 2, (l-n)].

Examination of the weights of a linear SVM classifier trained to discriminate drug from placebo sessions across all analgesic studies identified a pattern of positive and negative

weightings driving predictions. Negative weights, which would tend to correspond to weaker (or more negative) responses in the drug condition, were present in established pain regions, including the insula and anterior cingulate cortex. The negative areas matched reductions in activation identified in a mixed-effects GLM meta-analysis across all studies (Fig. 2B), and patterns of the effects of analgesics reported in past studies *(3, 19, 20)*. These regions, previously associated with analgesia and pain perception, are likely to have a variety of functions reflecting the multidimensional pain experience, along with attention, mood and salience *(30)*. There were also areas with positive weights, including the default mode network and temporal and occipital cortices (Fig. 2C).

Using a single established analgesic provides a less reliable test for evidence of analgesic clinical efficacy

To determine whether approaches using existing compounds could be effective when only a limited number of studies of efficacious compounds were available, we investigated the performance of algorithms trained on just one study to identify other analgesics (Fig. 3). These assessments were less reliable, producing a maximum mean accuracy of 68% (Naproxen) compared to a mean accuracy of 74% when classifiers were trained on all remaining studies. Assessments using the THC study as training data performed particularly poorly, matching the poor performance of the clinical efficacy assessment when applied to this study. Transfer of discrimination capabilities across studies was good: for both compounds tested in two studies (pregabalin, remifentanil), training on one of the two studies, and testing on the second, produced high drug-placebo discrimination rates (minimum accuracy 74%, $p < 0.02$) (Fig. 3).

Testing for clinical efficacy using signatures derived from modulation of stimulus intensity

We tested a second approach to clinical efficacy assessment, which used fMRI signatures generated from the same study, associated with a lessening of clinical symptoms. For instance, a classifier was trained to distinguish responses to allodynic stimuli from the less painful non-allodynic stimuli. This classifier was tested for its ability to distinguish responses to allodynic stimuli measured in drug and placebo sessions. This approach was less reliable than the multi-study strategy, although successful discrimination was achieved in studies showing strong pharmacodynamic effects [Table 3, (a,e,f)]. For several studies [Table 3, (d and h)], responses to the different stimulus levels could themselves not be reliably distinguished, rendering this approach impracticable. Some studies involved additional somatosensory stimuli that were less clinically relevant, such as mild thermal pain or pressure to non-allodynic regions of the skin. Drug-related modulations were not as reliably detected when using these stimuli, with no significant effect observed in four of eight assessments for pharmacodynamic effect (table S1).

Relation of imaging-based assessments to subject-reported pain ratings

Seven of the fMRI studies also measured pain self-reports [Table 1, (b to h)], with five identifying a significant average reduction in ratings of painful stimuli in the analgesic sessions (studies b, d,e,g, h). Two studies involving pregabalin and remifentanil [Table 1, (c and f)]*(21, 25)* showed evidence for drug action in the imaging data but not in the behavioral measures. Although the stimuli and pain-rating procedures used in these studies were not always optimized to detect changes in clinical pain, these results suggest that fMRI may be capable of providing information beyond what is available from behavioral measures.

Image-based was superior to co-ordinate based mega-analyses and region-of-interest analyses

We used an image-based mega-analytic approach to identify evidence of clinical efficacy *(20)*. We compared this to the widely used co-ordinate–based meta-analytic methods, which combine peak-activation co-ordinates across research reports to identify reliably activating regions (*20*, *29*). These approaches have the benefit of using information that is typically avaialble in research reports, but lack sensitivity for effects that are modest and spatially extended. Here, there were no areas where significant effects were observed in more than five of the eight studies (paired t-test, cluster-corrected for multiple comparisons; table S2), so this approach identified no consistent drug effects. Assessments using signals derived from a set of anatomically defined pain-related regions-of-interest (ROIs) were also less reliable than the image-based assessments using MVPA, with the procedure failing for pregabalin and tramadol studies (table S2)

DISCUSSION

Early-stage testing of candidate CNS compounds has often been inadequate. Accomplishing the 'Three Pillars of Survival'—exposure at the site of action, target binding and expression of functional pharmacological activity—has been associated with greater likelihood of candidate success *(2)*. fMRI has the potential to non-invasively and cost-effectively identify pharmacological activity (the third pillar), but integration of imaging methodologies into drug development decision-making is a challenge *(5, 17)*. Here, we have demonstrated and validated an fMRI-based assessment protocol, using published datasets, that was able to identify moderate to strong evidence for drug effects for all clinically proven analgesics, with no false-positives for control data. These results suggest that the fMRI can be of immediate value to analgesic drug

discovery. Our protocol provides a framework for the refinement of experimental and analytic methods. The framework can used with existing data sets, making it useful for rapid, early-stage drug assessment.

With its ability to characterize activity across multiple brain networks, fMRI has the potential to provide predictors for drug candidates targeting a variety of aspects of clinical efficacy, *(2, 3)*. We have demonstrated an approach that identifies predictors from sets of compounds with established efficacy, using the fact that the efficacy of these compounds has been validated through clinical trials. Even if some of the neural processes underlying response commonalities across these compounds may not be directly associated with disease symptoms, these secondary features can still be predictive of efficacy.

One concern with assessments using signatures derived from established compounds is that they will overlook compounds with novel or different modes of action. This is a possibility for any biomarker not directly linked to basic clinical outcomes. If a candidate compound is expected to have a substantially different mode of action, it would be prudent to include alternative efficacy assessments, for example approaches using correlates of clinical symptoms, such as pain. Our approach using signatures obtained by identifying correlates of variations in painful stimulus intensities performed poorly in some studies (table S1). For at least one compound, the assessment failed because there were no detectable differences in responses to the different stimulus levels in the training data. An alternative approach could be to identify a general marker of pain levels across multiple studies. This marker could potentially be a more reliable marker of efficacy than pain reports in populations where pain reports are unreliable (19). Defining and testing multiple approaches to identifying evidence for efficacious action will maximize the information that can be obtained from brain-imaging studies.

The prospects of fMRI-based assessments will vary across drug classes, and should be carefully assessed during experimental design. When an fMRI assessment protocol is implemented, decision rules will need to take into account the level of confidence in the test compound and the requirements of the specific drug discovery context. Decision rules might focus on rejecting compounds showing no evidence of pharmacodynamic effects, or accelerating compounds showing promising effects *(1)*. The potential for false-positives should be considered, as brain areas responsive to pain and analgesia are associated with a variety of other functions that will be modulated by numerous compounds *(30, 31)*. Drugs may modify a variety of physiological variables that can confound imaging signals.

Owing to the difficulty of obtaining good control datasets, the present assessment relied on the assumption that non-efficacious compounds will not consistently show effects similar to efficacious compounds. Incorporating non-efficacious compounds and other modulations of stimuli into classifier training will increase the specificity of these signatures, and enable better testing of protocols. Placebo effects may also present confounds to drug assessment, as they can reflect active brain processes providing symptom relief and may overlap and interact with the processes affected by drugs. As drugs were contrasted against placebo in the present work our results do not preclude the possibility that placebo effects were present. Approaches that can detect and account for placebo effects will enhance assessments.

The use of multi-study protocols for drug assessment and other applications will inform experimental design choices. For example, we found that clinically relevant stimuli enabled more reliable identification of drug effects. Additionally, we found that acute dosage studies tended to be more sensitive than studies where subjects received the compound for an extended period prior to scanning. fMRI scanning for acute-dosage studies was undertaken at peak drug

concentrations. In contrast, chronic-dosage studies involved scanning at lower steady state concentrations of drugs and involved patients with clinical pain conditions that are less reliably simulated by experimental stimuli. Altered brain responses under drug conditions may also normalize over extended dosage periods. MVPA methods need to be refined for the multi-study context. Approaches that into account the covariance structure of samples, variations in SNR across subjects and studies, and multiple modes of efficacious drug action, are likely to be beneficial. Classifiers that provide probabilistic estimates of the presence of a drug will be valuable, and potentially capable of identifying drug effects in smaller samples.

Obtaining the data necessary for image-based multi-study analyses can be a challenge, as these data are not typically made available when results are published. The present work required the collection of studies from several laboratories, and extensive vetting and standardization of data formats. However, the sensitivity achieved by image-based MVPA methods means that the multi-study approach can be implemented using smaller sets of studies. Crucially, they could be applied to a series of studies from a single research group (19, 26). The collation of datasets will also be aided by new tools that automate the generation and transfer of raw and processed data (*27*).

Our protocol could be used with other CNS imaging modalities, such as positron emission spectroscopy (PET), and electroencephalography (EEG), and even related physiological and behavioral measures. The protocol will be relevant whenever neural or behavioral correlates of the disease are non-specific or complex, and when complex multivariate or multi-modal data needs to be translated into validated decision procedures. It may be possible to identify correspondences between fMRI and other neuroimaging techniques that directly measure brain metabolite concentrations or receptor occupancies. This strategy has recently

been demonstrated in a study targeting reward-related mechanisms relevant to addiction, which characterized the effects of two opioid antagonists using [11C]-carfentanil PET and fMRI *(5)*.

Beyond drug development, the quantitative integration of existing data into analyses of new imaging studies can expand the range of inferences that can be made and help to structure and refine neuroimaging-based research. Multi-study approaches are important for many potential translational applications, such as prediction of drug response, disease state or disease progression, where extensive validation of biomarkers is essential to ensure sensitivity and specificity *(15, 19)*. These strategies can contribute to the identification and stratification of disease mechanisms, particularly in areas such as neuropsychiatry where the underlying pathophysiological processes are unclear and optimal imaging paradigms remain under investigation.

Materials and Methods

Study design

To determine whether fMRI can provide informative assessments of drug candidates, we tested our protocol on eight investigations of analgesics. We focused on analgesics because studies of analgesics are relatively common, enabling us to collate enough datasets published in the literature and on www.ClinicalTrials.gov to test multi-study methods. Analgesics are relevant to a variety of disease conditions and are a major target of drug discovery efforts. Our primary endpoint was the proportion of test analgesic datasets for which the proof-of-concept drug assessment procedure found evidence for analgesic drug action at the predefined rate, along with the proportion of false positives. We expected that for the protocol to be useful, at least 6 of 8

compounds should be identified as showing some evidence for drug effects, with no more than one false positive in the control data. Analysis methods were specified prior to any data analysis. Data acquisition and pre-processing of the individual studies were blinded. Unblinding of condition labels was necessary to generate cross-validated prediction accuracies.

Analgesic study datasets

We identified seven placebo-controlled randomized crossover fMRI studies investigating the effects of analgesics on pain responses (Table 1)*(6, 9, 18, 21–25)*. One study assessed two compounds (Table 1, c and d). Overall, the studies investigated the effects of six distinct analgesics on pain responses (two compounds were studied twice). Sample sizes of the individual studies ranged between 12 and 23 subjects, depending on specific experimental requirements. Studies were designed and acquired independently at the collaborating research groups and were combined after an agreement to share datasets.

The studies (Table 1) consisted of: (a) a study of gabapentin's effects on painful stimuli applied to hyperalgesic (capsaicin induced) and non-allodynic body surfaces directed by I. T. and the PAIN group at FMRIB Centre, University of Oxford *(9)*; (b) a comparison of pregabalin and placebo in fibromyalgia patients directed by R. E. H. and co-workers at the University of Michigan Chronic Pain and Fatigue Research Centre *(24)*; (c, d) a comparison of pregabalin, tramadol, and placebo for post-traumatic neuropathic pain (PTNP) directed by I. T. and the PAIN group at FMRIB (data published online under http://clinicaltrials.gov/show/NCT00610155*(25)*; information provided in Methods); (e) a study of remifentanil, assessing punctate and thermal stimuli directed by I. T. and the PAIN group at FMRIB *(6, 18)*; (f) a study of the effects of remifentanil on LASER pain and a number of non-

painful stimuli run by R. W. and co-workers at FMRIB *(21)*; (g) a study assessing the effects of tetrahydrocannabinol (THC) on painful stimuli to a region sensitized with capsaicin directed by I. T. at FMRIB *(23)*; (h) a comparison of naproxen and placebo for osteoarthritis by M. A. H. and colleagues at the Institute of Psychiatry, Kings College London*(22)*. Studies were performed according to relevant institutional guidelines. Informed consent was obtained from all study participants.

Four studies investigated subjects diagnosed with clinical conditions presenting with pain as the predominant symptom (Table 1, studies b to d, and h*)*. These studies used daily dosing lasting a period of 1 to 2 weeks prior to the scanning session. The other four studies (Table 1; a, e to g) administered analgesics to healthy subjects immediately prior to the imaging session. Pain reports were recorded in some studies, and could include rating of the pain intensity of individual trials, sessions, or of ongoing clinical pain. This variability in measurements is typical for imaging studies, where there are many experimental design options and relatively few studies.

Several assessments of control data with no analgesic effect were performed to determine how the procedure would perform with inadequate data or ineffective drugs (Table 1). Studies that assessed the effects of drugs that do not reduce pain were not available as such studies are rare.

The datasets were pre-processed and analyzed using GLMs, described in Supplementary Methods, to produce trial, session and study-level parameter and variance maps reflecting responses to painful stimuli during placebo or drug conditions. These maps were the primary inputs to the drug-assessment procedure.

Additional experimental methods for pregabalin and tramadol experiment

This experiment investigated the effects of pregabalin (study c) and tramadol (study d) on brain responses to painful stimuli in patients with post-traumatic neuropathic pain (PTNP) (Table 1). It was a double-blinded crossover study with pregabalin (titrated to 150 mg BID) tramadol SR (titrated to 200 mg BID), and placebo, in randomized order. The study was approved by the Oxford Research Ethics Committee C (08/H0606/5) and registered with the ClinicalTrials.gov (NCT00610155), where study details, behavioral outcomes, and ROI-based FMRI outcomes are provided (http://clinicaltrials.gov/show/NCT00610155).

The study assessed twenty-one patients with PTNP, with 16 completing the study. Inclusion criteria for the study included: 1). a diagnosis of neuropathic pain for a duration of at least three months; 2). dynamic mechanical allodynia (DMA) with an intensity of no less than 4 on the 11 point Numerical Rating Scale (NRS); 3). an average NRS pain intensity of at least 3 over the previous week. The site of DMA varied across subjects, with lateralization balanced across the study.

Imaging sessions occurred at the end of each 7 day dosage period, and were followed by a 7 day washout period. Imaging sessions included FMRI scans recording brain responses to the delivery of 15 Somedic brush stimuli to affected areas (DMAa) and control areas (DMAc). These brush stimuli lasted approximately five seconds, with a mean inter-stimulus period of 40 seconds. Subjects were asked to score the average pain intensity of each stimulus type, which were given in separate blocks. An 11-point NRS was used, with 0 corresponding to “no pain” and 10 corresponding to “worst pain possible”. A present Pain Intensity (PPI) score (ongoing background pain at the beginning of the scanning session) was also recorded. BOLD fMRI data acquisition is described in Supplementary Methods.

Validation of prediction methods

To determine the value of the multi-study approach for reliable signatures of efficacy, we compared it to how well algorithms trained on individual studies could identify analgesics in other studies. To assess the value of the multivariate prediction strategy, we examined the ability of ROIs to identify efficacious action. The ROIs consisted of the posterior, medial, and anterior insula cortices; anterior and posterior cingulate cortices; nucleus accumbens; nucleus cuneiformis; amygdala; caudate; hippocampus; precuneus cortex; putamen; rostral ventromedial medulla; primary and secondary sensory cortices; and sensory thalamus. The ROIs were defined from T1 MR anatomical images, by experts in functional brain anatomy (K. W., I. T.). The ROI data were analyzed in an identical manner to the ICA components derived from the whole-brain data.

Statistical analysis

The drug-action assessment produced a study prediction accuracy rate through a leave-one-subject-out cross-validation procedure. As the algorithm for the drug efficacy assessment was trained on other studies, the same trained algorithm was applied to all subjects. Probabilities that the resulting discrimination accuracies for each study were produced by chance were calculated (binomial test), along with 90% Wilson-score confidence intervals for the mean accuracy. Owing to the differences in these tests, a study can show a confidence interval encompassing 0.5, while still being $p<0.05$ for the hypothesis that discrimination was at chance level.

The multi-level GLM analyses of the BOLD fMRI data produced T-statistic images reflecting the voxel-wise statistical strength of the stimulus-related BOLD signal responses to painful stimuli (one-sided, one-group t-test) or the effects of drugs on these responses (drug – placebo, within subject: paired one-sided, two-group t-test). These analyses used temporal autocorrelation correction and outlier detection to ensure model validity *(32, 33)*. The resulting maps were transformed into Z-statistic images and thresholded (correcting for multiple comparisons) using FSL FEAT *(33)*. A cluster forming threshold of Z=2.3 was first applied. Then, a cluster extent threshold was defined, based on Gaussian Random Field theory, identifying clusters with a significance level of $p<0.05$ *(34)*.

Supplementary Materials

Supplementary Methods

Fig. S1. Alternative clinical efficacy procedure.

Fig. S2. Modeling of pain responses in the pregabalin study (b) *(24)*.

Fig. S3**.** Summary of significant effects in individual study placebo vs. drug pain response contrasts for individual analgesic studies in Table 2.

Table S1. Results of procedure applied to analgesic studies when experimental stimuli with lower clinical relevance are used.

Table S2. Outcome of the analgesic assessment protocol when using inputs derived from a set of pain-related ROIs.

Refs (*36–40*)

References and Notes:

1. R. G. Wise, C. Preston, What is the value of human FMRI in CNS drug development?, *Drug Discov. Today* **15**, 973–980 (2010).

2. P. Morgan, P. H. Van Der Graaf, J. Arrowsmith, D. E. Feltner, K. S. Drummond, C. D. Wegner, S. D. A. Street, Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival, *Drug Discov. Today* **17**, 419–424 (2012).

3. D. Borsook, L. Becerra, R. Hargreaves, Biomarkers for chronic pain and analgesia. Part 1: the need, reality, challenges, and solutions., *Discov. Med.* **11**, 197–207 (2011).

4. Y. Agid, G. Buzsaki, D. M. Diamond, R. Frackowiak, J. Giedd, J.-A. Girault, A. Grace, J. J. Lambert, H. Manji, H. Mayberg, M. Popoli, A. Prochiantz, G. Richter-Levin, P. Somogyi, M. Spedding, P. Svenningsson, D. Weinberger, How can drug discovery for psychiatric disorders be improved?, *Nat Rev Drug Discov* **6**, 189–201 (2007).

5. E. A. Rabiner, J. Beaver, A. Makwana, G. Searle, C. Long, P. J. Nathan, R. D. Newbould, J. Howard, S. R. Miller, M. A. Bush, S. Hill, R. Reiley, J. Passchier, R. N. Gunn, P. M. Matthews, E. T. Bullmore, Pharmacological differentiation of opioid receptor antagonists by molecular and functional imaging of target occupancy and food reward-related brain activation in humans., *Mol. Psychiatry* **16**, 826–35, 785 (2011).

6. V. Wanigasekera, M. C. H. Lee, R. Rogers, P. Hu, I. Tracey, Neural correlates of an injury-free model of central sensitization induced by opioid withdrawal in humans., *J. Neurosci.* **31**, 2835–42 (2011).

7. B. G. Oertel, C. Preibisch, T. Wallenhorst, T. Hummel, G. Geisslinger, H. Lanfermann, J. Lotsch, Differential Opioid Action on Sensory and Affective Cerebral Pain Processing, *Clin Pharmacol Ther* **83**, 577–588 (2007).

8. R. Rogers, R. G. Wise, D. J. Painter, S. E. Longe, I. Tracey, An investigation to dissociate the analgesic and anesthetic properties of ketamine using functional magnetic resonance imaging., *Anesthesiology* **100**, 292–301 (2004).

9. G. D. Iannetti, L. Zambreanu, R. G. Wise, T. J. Buchanan, J. P. Huggins, T. S. Smart, W. Vennart, I. Tracey, Pharmacological modulation of pain-related brain activity during normal and central sensitization states in humans, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18195–18200 (2005).

10. N. Khalili-Mahani, R. M. W. Zoethout, C. F. Beckmann, E. Baerends, M. L. de Kam, R. P. Soeter, A. Dahan, M. A. van Buchem, J. M. A. van Gerven, S. A. R. B. Rombouts, Effects of morphine and alcohol on functional brain connectivity during "resting state": a placebo-controlled crossover study in healthy young men., *Hum. Brain Mapp.* **33**, 1003–18 (2012).

11. J. Upadhyay, J. Anderson, R. Baumgartner, A. Coimbra, A. J. Schwarz, G. Pendse, D. Wallin, L. Nutile, J. Bishop, E. George, I. Elman, S. Sunkaraneni, G. Maier, S. Iyengar, J. L. Evelhoch, D. Bleakman, R. Hargreaves, L. Becerra, D. Borsook, Modulation of CNS pain circuitry by intravenous and sublingual doses of buprenorphine, *Neuroimage* **59**, 3762–3773 (2012).

12. G. A. Cecchi, L. Huang, J. A. Hashmi, M. Baliki, M. V Centeno, I. Rish, A. V. Apkarian, A. Morrison, Ed. Predictive dynamics of human pain perception., *PLoS Comput. Biol.* **8**, e1002719 (2012).

13. A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, J. Mourão-Miranda, Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes., *Neuroimage* **49**, 2178–2189 (2010).

14. K. Wiech, I. Tracey, The influence of negative emotions on pain: Behavioral effects and neural mechanisms, *Neuroimage* **47**, 987–994 (2009).

15. J. A. Hashmi, A. T. Baria, M. N. Baliki, L. Huang, T. J. Schnitzer, A. V. Apkarian, Brain networks predicting placebo analgesia in a clinical trial for chronic back pain., *Pain* **153**, 2393–402 (2012).

16. J. Upadhyay, J. Anderson, A. J. Schwarz, A. Coimbra, R. Baumgartner, G. Pendse, E. George, L. Nutile, D. Wallin, J. Bishop, S. Neni, G. Maier, S. Iyengar, J. L. Evelhoch, D. Bleakman, R. Hargreaves, L. Becerra, D. Borsook, Imaging Drugs with and without Clinical Analgesic Efficacy, *Neuropsychopharmacology* **36**, 2659–2673 (2011).

17. D. Borsook, J. Upadhyay, M. Klimas, A. J. Schwarz, A. Coimbra, R. Baumgartner, E. George, W. Z. Potter, T. Large, D. Bleakman, J. Evelhoch, S. Iyengar, L. Becerra, R. J. Hargreaves, Decision-making using fMRI in clinical drug development: revisiting NK-1 receptor antagonists for pain, *Drug Discov. Today* (2012), doi:10.1016/j.drudis.2012.05.004.

18. V. Wanigasekera, M. C. Lee, R. Rogers, Y. Kong, S. Leknes, J. Andersson, I. Tracey, Baseline reward circuitry activity and trait reward responsiveness predict expression of opioid analgesia in healthy subjects., *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17705–10 (2012).

19. T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, E. Kross, An fMRI-based neurologic signature of physical pain., *N. Engl. J. Med.* **368**, 1388–97 (2013).

20. G. Salimi-Khorshidi, S. M. Smith, J. R. Keltner, T. D. Wager, T. E. Nichols, Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies, *Neuroimage* **45**, 810–823 (2009).

21. R. G. Wise, S. Dirckx, E. Russell, A. Johnson, K. Pattinson, G. D. D. Iannetti, in *16th Annual Meeting of the Organization for Human Brain Mapping*, (2010), p. 2589.

22. S. Duncan, K. Kristina, J. O'Muircheartaigh, M. A. Thacker, J. P. Huggins, B. Vennart, N. J. Massat, E. Choy, S. C. R. Williams, M. A. Howard, Pharmacological Modulation of Hand Pain in Osteoarthritis: A Double-Blinded Placebo-Controlled functional Magnetic Resonance Imaging Study Using Naproxen, *Rev.* (2015).

23. M. C. Lee, M. Ploner, K. Wiech, U. Bingel, V. Wanigasekera, J. Brooks, D. K. Menon, I. Tracey, Amygdala activity contributes to the dissociative effect of cannabis on pain perception., *Pain* **154**, 124–34 (2013).

24. R. E. R. Harris, V. Napadow, J. P. Huggins, L. Pauer, J. Kim, J. Hampson, P. C. Sundgren, B. Foerster, M. Petrou, T. Schmidt-Wilcke, D. J. Clauw, Pregabalin Rectifies Aberrant Brain

Chemistry, Connectivity, and Functional Response in Chronic Pain Patients, *Anesthesiology* , 1453–64 (2013).

25. K. Wartolowska, J. Huggins, E. P. Duff, W. Vennart, P. Rogers, B. Hoggard, I. Tracey, A Methodology Study Of Brain Imaging Of Pain-Killers In Post-Traumatic Neuropathic Pain Patients*clinicaltrials.gov* (2013) (available at http://clinicaltrials.gov/ct2/show/NCT00610155).

26. F. Pedregosa, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, A. Gramfort, in *Machine Learning in Medical Imaging, Lecture Notes in Computer Science Volume 7588*, F. Wang, D. Shen, P. Yan, K. Suzuki, Eds. (Springer, 2012), pp. 234–241.

27. V. Vapnik, S. E. Golowich, A. Smola, in *Advances in Neural Information Processing Systems 9*, (1996), vol. 9, pp. 281–287.

28. F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview., *Neuroimage* **45**, S199–S209 (2009).

29. E. P. Duff, A. J. Trachtenberg, C. E. Mackay, M. A. Howard, F. Wilson, S. M. Smith, M. W. Woolrich, Task-driven ICA feature generation for accurate and interpretable prediction using FMRI, *Neuroimage* , 189–203 (2012).

30. V. Legrain, G. D. Iannetti, L. Plaghki, A. Mouraux, The pain matrix reloadedA salience detection system for the body, *Prog. Neurobiol.* **93**, 111–124 (2011).

31. K. D. Davis, E. Racine, B. Collett, Neuroethical issues related to the use of brain imaging: can we and should we use brain imaging as a biomarker to diagnose chronic pain?, *Pain* **153**, 1555–9 (2012).

32. M. Woolrich, Robust group analysis using outlier inference., *Neuroimage* **41**, 286–301 (2008).

33. M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, S. M. Smith, FSL., *Neuroimage* **62**, 782–90 (2012).

34. K. J. Worsley, in *Functional MRI - An Introduction to Methods*, P. Jezzard, P. M. Matthews, S. M. Smith, Eds. (2001), pp. 251–270.

35. M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, M. Jenkinson, S. M. Smith, Multilevel linear modelling for FMRI group analysis using Bayesian inference., *Neuroimage* **21**, 1732–47 (2004).

36. C. F. Beckmann, S. M. Smith, Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging, *IEEE Trans. Med. Imaging* **23**, 137–152 (2004).

37. M. Hanke, Y. O. Halchenko, P. B. Sederberg, E. Olivetti, I. Fründ, J. W. Rieger, C. S. Herrmann, J. V Haxby, S. J. Hanson, S. Pollmann, PyMVPA: A Unifying Approach to the

Analysis of Neuroscientific Data., *Front. Neuroinform.* **3** (2009), doi:10.3389/neuro.11.003.2009.

38. C. C. Chang, C. J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).

39. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in {P}ython, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgments: We thank E. Berry, T. Smart, J. Huggins, H. Cagnan, and K. Murphy for helpful discussions. **Funding:** Pfizer funded this study, and studies (b, c, d, and h) in Table 1. The UK Medical Research Council and Wellcome Trust provided core funding of the FMRIB center facilitating this work and studies (a, c, e, f, and g) in Table 1. M.W.W. and I.T. are funded by the Wellcome Trust, and supported by the NIHR Oxford Biomedical Research Centre based at Oxford University Hospitals Trust Oxford University. **Author contributions:** E.P.D. collated the datasets, analyzed the data, and wrote the manuscript. E.P.D., S.M.S., M.A.H., M.W.W., M.W., I.T., R.W., W.V., and F.J.W. developed and critiqued the methodological strategy and facilitated data collation. I.T. and W.V. designed and led study (a). R.E.H. designed and acquired data in study (b). I.T., K.W., and V.W. designed and acquired data for studies (c and d). V.W. and I.T. designed and acquired data for study (e). R.W. designed and acquired data for study (f). M.L. and I.T. designed and acquired data for study (g). M.A.H. designed and acquired data for study (h). All authors contributed the manuscript. **Competing interests:** W.V. was a full-time employee of Pfizer Ltd and now works as a contractor for them; he holds stock options in Pfizer Inc. M.W. is an employee of Pfizer Ltd; he holds stock options in Pfizer Inc. F.J.W. was an employee of Pfizer Ltd and held stock options in Pfizer Inc. before, during, and after the work reported, but not during manuscript preparation; he is currently a consultant to GlaxoSmithKline plc, IPPEC, King's College London, Lundbeck A/S, Mentis Cura, and Pfizer Inc. I.T. has been an ad hoc consultant, advisory board member, and educational lecturer with Pfizer, Merck, Grunenthal, and Lilly. R.H. has received consulting fees for Pfizer related to this work and received funding for the University of Michigan study. **Data and materials availability:** The software used to perform these analyses is available for free download (FSL: www.fmrib.ox.ac.uk/fsl; PyMVPA: www.pymvpa.org). A new function in FSL, bundleFEAT, has been designed to enable rapid collation and sharing of fMRI study analyses. . Individual clinical study data can be requested by contacting the respective PI for each publication. Raw data of the meta-analytic analysis (the inputs into the classifiers and associated meta-data) can be obtained through an MTA by contacting EPD.

Figure captions

Fig. 1. Protocol for assessing candidate CNS compounds using fMRI. Top row shows the general protocol. Bottom row shows a schematic of the proof-of-concept analgesic assessment protocol. (**A**) Quality-assurance. Prior to the assessment of drug effects, quality-assurance

procedures test that the experimental protocol, preprocessing, and modeling identify expected brain responses. A comparison was performed between the effects of the painful stimulus in the test study and effects consistently observed in studies using similar stimuli. A study with lower responses in relevant brain areas was flagged as potentially problematic (t-test, corrected for multiple comparisons). (**B**) Test for pharmacodynamic effect. This stage tests whether the compound reliably modulates brain activity. Consistent modulation of brain responses is evidence for a pharmacodynamic effect. In the proof-of-concept procedure, a pharmacodynamic effect was flagged if an algorithm could be trained to distinguish drug from placebo responses in held-out subjects. **(C)** Test for evidence of clinical efficacy. The final stage tests whether observed modulations of responses correspond to established signs of efficacious action. The proof-of-concept approach determined whether a classifier trained to detect the effects of established analgesics could identify the test compound.

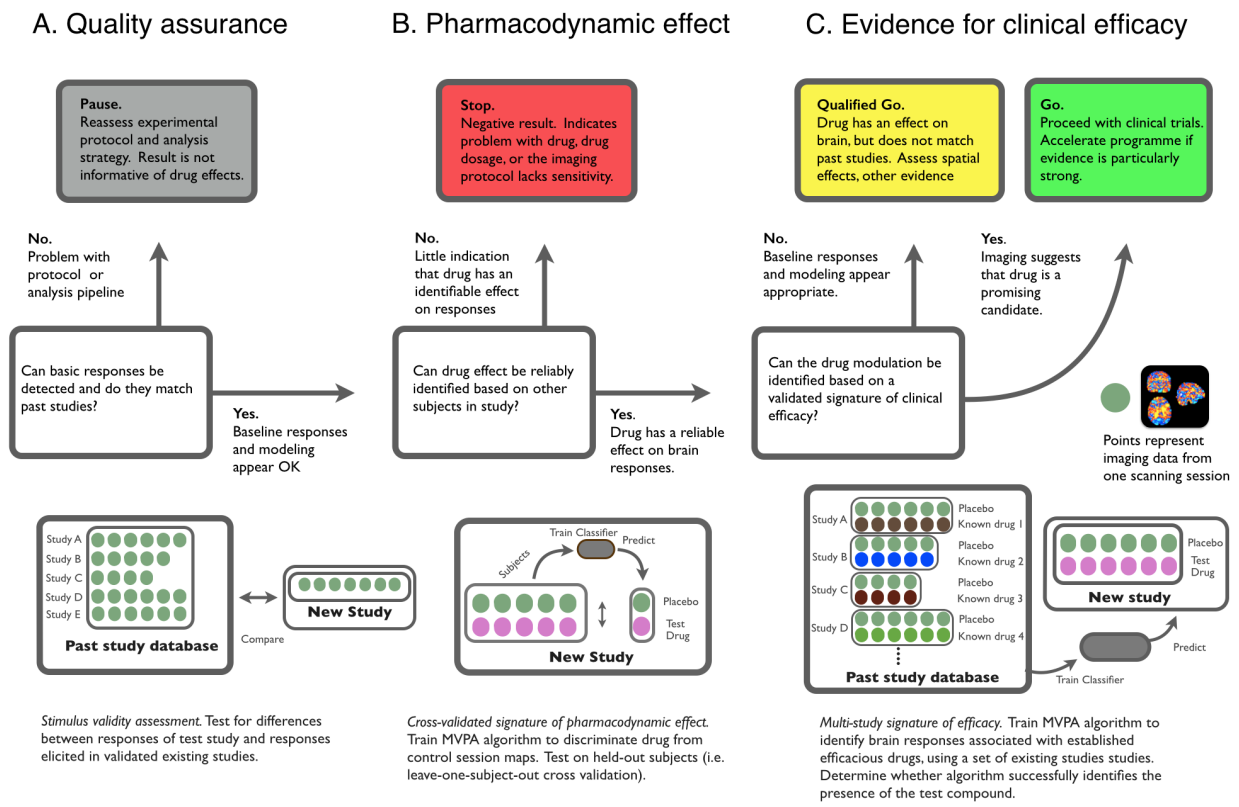

Fig. 2. Multi-study effect maps. (**A**) Thresholded z-statistic maps of a mixed effects analysis of the average pain response across all 8 studies in Table 1. Red corresponds to brain regions showing significant response to the painful stimuli (one-group t-test for nonzero response, $p<0.05$, cluster-corrected). (**B**) Mixed effects analysis of modulation of pain responses by analgesics across all 8 analgesic studies in Table 1. Blue corresponds to regions showing a significantly reduced response to the painful stimuli in the analgesic sessions (paired two-group t-test, $p<0.05$, cluster-corrected). (**C**) Weight map of the SVM trained across all 8 studies in Table 1 (arbitrary threshold). Red corresponds to positive weights; blue, to negative weights.

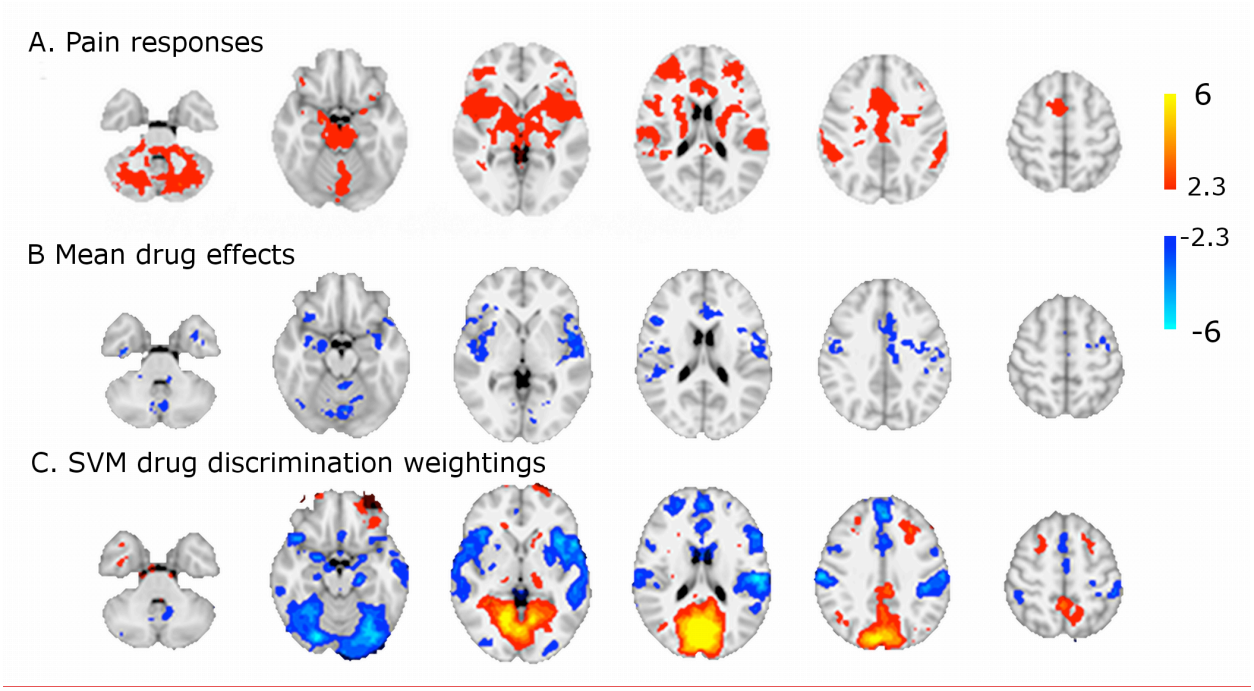

Fig. 3. Identifying evidence for clinical efficacy using data from a single training study. Classifiers trained to discriminate a single analgesic from placebo were tested on other studies to determine the extent to which discriminative capabilities transferred. "All" corresponds to the training set comprising of all other studies, excluding studies of the same compound. The diagonal corresponds to within-study prediction rates measured using leave-one-subject-out cross-validation. The 'All' row shows the prediction accuracies for predictors trained on multiple studies (Fig. 2). The "mean" column shows the average performance of each training dataset when applied to other studies. Black boxes indicate discrimination was not greater than 0.5.

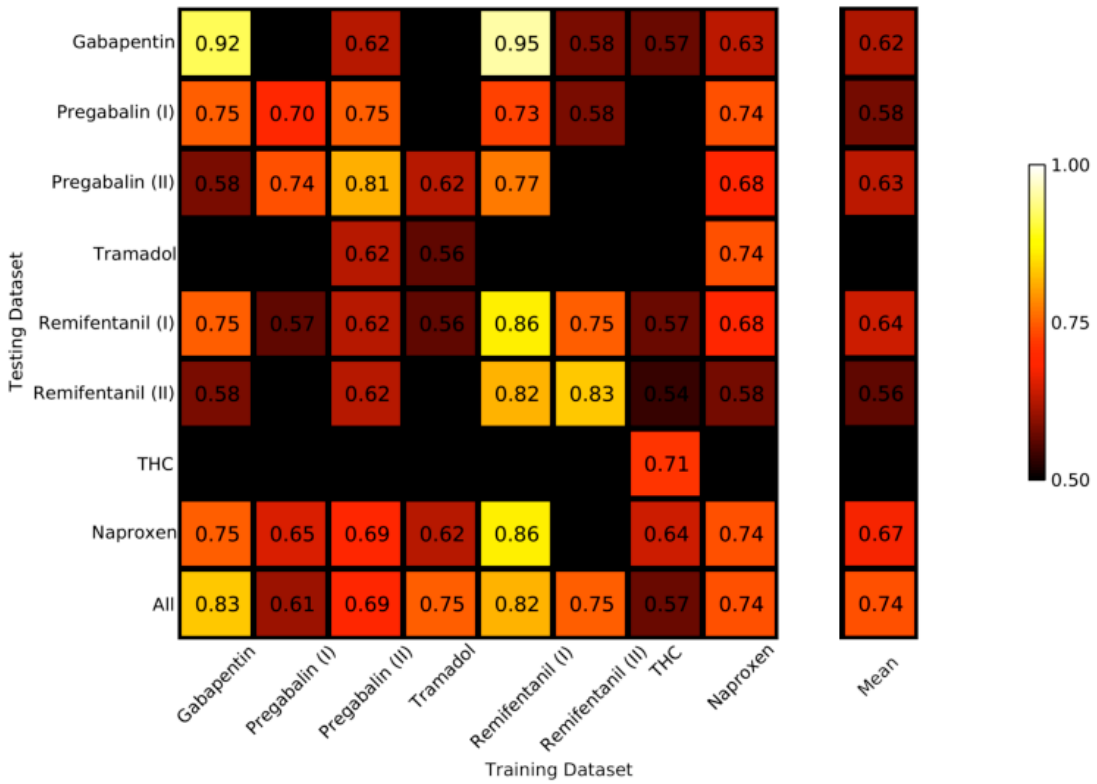
Gabapentin
Pregabalin (I)
Pregabalin (II)
Tramadol
Remifentanil (I)
Remifentanil (II)
THC
Naproxen
All
Testing Dataset
0.92 0.62 0.95 0.58 0.57 0.63
0.75 0.70 0.75 0.73 0.58 0.74
0.58 0.74 0.81 0.62 0.77 0.68
0.62 0.56 0.74
0.75 0.57 0.62 0.56 0.86 0.75 0.57 0.68
0.58 0.62 0.82 0.83 0.54 0.58
0.71
0.75 0.65 0.69 0.62 0.86 0.64 0.74
0.83 0.61 0.69 0.75 0.82 0.75 0.57 0.74
Gabapentin
Pregabalin (I)
Pregabalin (II)
Tramadol
Remifentanil (I)
Remifentanil (II)
THC
Naproxen
Training Dataset
0.62
0.58
0.63
0.64
0.56
0.67
0.74
Mean
1.00
0.75
0.50

Table 1. Analgesic and control study experimental parameters. Overview of 8 studies used in the testing and validation of our drug assessment procedure. Studies (c) and (d) were acquired in the same experiment and were separated in all assessments. PTNP, post-traumatic neuropathic pain; N/R, not recorded; N/A, not applicable; BID, *bis in diem* (in two divided doses per day); BPC, blood plasma concentration; VAS, visual analogue scale.

Study	Drug (reference/ clinicaltrials.gov id)	Patient condition	*n* subjects	Scanner	Dose	Stimuli	*n* trials	Pain score
				Analgesic drug study assessments				
a	Gabapentin *(9)*	Healthy	12	3T Varian	1800 mg oral (taken 2 h prior)	Punctate to hyperalgesic skin	20	N/R
b	Pregabalin *(24)*	Fibromyalgia	23	3T GE	225 mg/day oral (7 days daily dosoing)	Thumb squeeze	6	Yes
c	Pregabalin *(25)* (NCT00610155)	PTNP	16	3T TIM Trio	150mg oral BID (7 days daily dosing)	Brush-evoked allodynia	15	No
d	Tramadol *(25)* (NCT00610155)	PTNP	16	3T TIM Trio	200mg oral BID (7 days daily dosing)	Brush evoked allodynia	15	Yes
e	Remifentanil *(6, 18)*	Healthy	22	3T Varian	2 ng/ml BPC i.v.	Punctate and thermal	10	Yes
f	Remifentanil *(21)*	Healthy	12	3T Varian	1.5 ng/ml BPC i.v	Laser	50	No
g	Tetrahydrocannabinol *(23)*	Healthy	14	3T Varian	15 mg oral (taken 2 h prior)	Punctate to hyperalgesic skin	20	Yes
h	Naproxen*(22)*	Osteoarthritis	19	3T GE HDx	220 mg oral (taken 1 h prior)	Key turn	15	Yes
				Control study assessments				
i	2nd Placebo (Study b)	Fibromyalgia	23	3T GE	N/A	Thumb squeeze	6	No
j	2nd Placebo (Study e)	Healthy	22	3T Varian	N/A	Punctate and thermal	10	No
k	2nd Placebo (Study f)	Healthy	12	3T Varian	N/A	LASER	50	No
l	Remifentanil (Study f)	Healthy	12	3T Varian	1.5 ng/ml BPC infusion	Flash	50	N/A
m	Remifentanil (Study f)	Healthy	12	3T GE	1.5 ng/ml BPC infusion	Brief tone	50	N/A

n	Naproxen (Study h)	Osteoarthritis	19	3T GE HDx	220 mg (1 h)	Visual stimulus	15	N/A

Table 2. Results of the proof-of-concept drug assessment protocol applied to analgesic and control studies. QA identified the percentage of voxels showing responses that were significantly lower than responses in regions consistently responding to similar stimuli in the other studies (unpaired t-test, corrected for multiple comparisons). The pharmacodynamic effect assessment determined whether the stimulus responses associated with the test compound could be distinguished from those associated with placebo, evidence of a drug effect. The clinical efficacy assessment tested whether studies of established compounds identified patterns of effects that could be used to distinguish the test compound from placebo. Accuracies give the proportion of subjects for which the test compound session was correctly identified. P-values indicate the probability of achieving this accuracy or better given no drug effect (binomial test, chance = 50%). Sample sizes correspond to the number of subjects reported in Table 1. 90% Wilson-score confidence intervals are also shown. Colors indicate whether compounds passed a specific assessment phase based on the predefined decision rules (Fig. 1): green – pass; yellow – provisional pass; red – fail; gray – problem with QA, reassess.

Table 2

Study	Test compound	QA	Pharmacodynamic effect		Clinical efficacy		Decision
		Area of reduced response (%)	Accuracy (range)	*P*	Accuracy (range)	*P*	
			Analgesic drug study assessments				
a	Gabapentin	0	92% (70-100)	0.0002	83% (60-94)	0.003	Go
b	Pregabalin	0	70% (52-83)	0.017	61% (44-76)	0.105	Go
c	Pregabalin	0	81% (61-92)	0.002	69% (42-79)	0.038	Go
d	Tramadol	0	56% (36-74)	0.22	75% (55-74)	0.01	Go (Q)
e	Remifentanil	0	86% (70-94)	0.000	82% (65-92)	0.000	Go
f	Remifentanil	0	83% (60-94)	0.003	75% (51-90)	0.003	Go
g	Tetrahydrocannabinol (THC)	0	71% (49-86)	0.028	57% (36-76)	0.22	Go (Q)
h	Naproxen	0	73% (55-83)	0.01	73% (55-83)	0.01	Go
			Control study assessments				
i	Placebo (Study b)	0	52% (22-54)	0.339	48% (32-64)	0.5	Stop
j	Placebo (Study e)	0	27% (15-45)	0.97	45% (29-62)	0.584	Stop
k	Baseline (Study f)	0	58% (36-78)	0.194	33% (16-57)	0.806	Stop
l	Visual stimulus (Study f)	79	74% (51-90)	0.019	33% (16-57)	0.019	Reassess
m	Auditory stimulus (Study f)	6	100% (82-100)	0.000	58%(36-78)	0.194	Reassess
n	Visual stimulus (Study h)	48	73% (55-87)	0.009	47% (30-65)	0.5	Reassess

Table 3. **Comparison of efficacy assessments.** Comparison of the clinical efficacy assessment shown in Table 2 with an alternative clinical efficacy assessment not requiring multi-study data. This assessment tested for a "normalization" of responses to disease-related stimuli in the drug condition. A classifier was trained to discriminate between the effects of different levels of a disease-related stimulus (e.g. different levels of alloydynia). Then, this classifier was tested for its ability to discriminate responses from drug and placebo sessions. Evidence for efficacy was identified if responses in the drug condition were consistently identified as responses associated with lower stimulus intensity. P values indicate the probability of achieving the obtained accuracy or better given no drug effect (chance = 50%, Binomial test). 90% Wilson-score confidence intervals are shown. Colors indicate whether compounds passed a specific assessment phase based: Green –pass; Yellow – qualified pass (Q); Red –fail.

Study	Drug	Clinical efficacy A (Existing analgesics)		Clinical efficacy B (Response normalization)			
		Accuracy (range)	*P*	Training comparison	Training accuracy (range)	Validation accuracy (range)	*P*
Analgesic study assessments							
a	Gabapentin	83% (60-94)	0.003	Hyperalgesic vs. normal	92% (70-100)	67% (43-84)	0.007
b	Pregabalin	61% (44-76)	0.105	High vs. low pain	70% (52-83)	39% (24-56)	0.800
c	Pregabalin	69% (42-79)	0.038	Allodynic vs. normal	69% (48-84)	63% (42-79)	0.105
d	Tramadol	75% (55-74)	0.01	Allodynic vs. normal	50% (31-69)	44% (26-64)	0.600
e	Remifentanil	82% (65-92)	0.000	Allodynic vs. normal	91% (76-97)	82% (65-92)	0.000
f	Remifentanil	75% (51-90)	0.003	Pain vs. non-pain	83% (60-94)	75% (51-90)	0.003
g	THC	57% (36-76)	0.22	Hyperalgesic vs. normal	71% (49-86)	57% (36-76)	0.210
h	Naproxen	73% (55-83)	0.01	High vs. low pain	63% (44-79)	58% (39-74)	0.180
Control study assessments							
i	Placebo (Study b)	52% (22-54)	0.339	High vs. low pain	70% (52-83)	30% (17-48)	0.890
j	Placebo (Study e)	27% (15-45)	0.97	Allodynic vs. normal	91% (76-97)	19% (33-67)	0.42

MATERIALS AND METHODS

Additional experimental methods for pregabalin and tramadol experiment

A 3-T Tim Trio scanner (Siemens) was used. BOLD fMRI was acquisitions used an EPI sequence with GRAPPA acceleration factor 2, a 192x192mm FOV and matrix size of 64x64. The sequence had a repetition time of 3 s, a TE of 30 ms and 87 degree flip angle. A symmetric-asymmetric spin echo sequence was used to acquire field maps to correct for distortions due to B0-field inhomogeneities: TR=532 ms, TE1= 5.19 ms, TE2=7.65 ms, flip angle=60 degrees. High resolution structural scans were acquired using an MPRAGE sequence: single shot, FOV=192 mm, matrix size 192x174 with a 1x1x1 mm voxel size, TR=2040 ms, TE=4.7 ms, IR=900 ms, flip angle 8 degrees.

Image pre-processing and General Linear Modeling (GLM) analysis for all studies

Datasets including raw time series images were transferred to the FMRIB Centre, Oxford University, where they were stored, analyzed, and assessed for quality in a standardized manner. Each dataset included functional MRI scans and a high-resolution structural scan. Two-level (within- and between-subjects) fMRI GLM analyses were performed for each study to produce the inputs used by the assessment protocol. The key inputs for the assessment protocol were response maps for individual subject's placebo and drug sessions, and study-level drug-effect maps. The multi-study analysis methods were specified prior to any data analysis. Acquisition and pre-processing of the individual studies was performed in a blinded manner. Unblinding was required for the cross-validated analyses.

Standard preprocessing and mapping analysis was employed using tools from FMRIB's Software Library (FSL) package (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Preprocessing included brain extraction (BET), head motion correction (McFLIRT), and spatial smoothing (SUSAN) (33). Appropriate controls for the effects of non-neural physiological processes, such as changes in respiration, heart rate, and head movement, were important, as these could induce false-positives or -negatives. Scans were excluded if an excess of 3mm of motion was detected across frames. Magnetic field maps were used to reduce distortions due to field inhomogeneities when available (FUGUE).

For each study, the pre-processed functional data from the placebo and drug sessions were fit with a temporal model composed of separate regressors for each of the different painful stimulus conditions employed in the study, along with additional regressors modeling non pain-related signals, such as innocuous sensory stimuli, pain-rating periods, physiological measurements, and head motion. The amplitude of the regressors was kept consistent across all studies to make them comparable. The FSL FMRI Expert Analysis Tool (FEAT) was used for linear modeling (33, 35). Resulting parameter maps were registered to standard anatomical space, via their high-resolution structural images using the FSL linear and non-linear registration tools FLIRT and FNIRT (33). These drug and placebo response maps were used in the classification assessments, and also fed into the higher-level study and multi-study GLM analyses.

Study-level effect maps were generated from session-level response maps using a second-level paired GLM that modeled consistent pair-wise differences between drug and placebo sessions. For studies with multiple types of stimuli, the stimuli expected to be most strongly modulated by analgesics were used. The study-level models included a regressor

modeling the average difference between drug and placebo session responses across all subjects. Additional regressors modeled the average stimulus response for each subject, across both the drug and placebo sessions. A further regressor modeled any effects associated with the order in which sessions were acquired. Separate models were used to generate the average response to the stimulus. These models were estimated using FLAME (FMRIB's Local Analysis of Mixed Effects), which provides a Bayesian mixed effects analysis incorporating estimates of uncertainty of the measured responses in each study (35).

The image-based meta-analyses were performed in the same manner as the higher level group analyses, utilizing parameter and variance maps from the individual studies. Regressor amplitudes were standardized across studies. No further normalization of maps was performed; this remains an area open to optimization. At the multi-study level we modeled mean effects. In the quality assurance assessment, we used a model accounting for mean differences between the pain response in a target study and all other studies.

Quality assurance (QA)

Meta-analytic approaches rely on the integrity of the datasets. In addition to standard quality control procedures built into the FSL software tools (33), we implemented an additional, meta-analytic assessment designed to identify modeling issues, such as inadequate baseline brain responses, incorrect timing specification, inaccurate hemodynamic models, or non-optimal filtering, that are unlikely to be identified from low-level data assessment. This assessment determined whether the baseline patterns of brain activity of the test study (e.g. pain responses) deviated significantly from those observed in past studies. Deviations will indicate that the data may not provide a valid assessment for drug effects. Ideally this activity assessment would be

performed on baseline sessions not affected by the drug or placebo arms of the study. However, these were generally not available, so the assessment was performed using both placebo and drug arms of the studies to ensure that no bias towards one condition is introduced.

The assessment focused on regions that were significantly activated or deactivated across the other studies in a GLM-based meta-analysis performed using the FSL FEAT software (33). A study was flagged as potentially problematic if it showed significantly lower responses compared to the remaining studies, in regions found to be consistently activated across all studies (significant regions here identified by the spatial thresholding described in the main text under "Statistical analysis").

Flagged studies were subject to further assessments to determine the source of the unexpectedly low values. Model fits were compared to average responses, to identify modeling errors such as incorrect timing specification, inaccurate HRF models, or non-optimal filtering. If other stimuli were used, these were investigated to determine whether responses to these stimuli were affected. Any problems that were identified were corrected, and all analyses were repeated with the corrected data.

Multivariate pattern analysis

Multivariate pattern analysis algorithms were trained to identify sessions with drug effects from held out single-subject data. Single-subject drug and placebo trial-wise regression coefficient maps from the individual subject GLM analyses were input into the prediction algorithms. Regressor amplitudes were standardized across studies. As with the GLM meta-analyses, no further normalization of effect maps was carried out.

Prior to prediction, data were projected onto a spatial basis set derived from an Independent Component Analysis (ICA) decomposition performed using the FSL tool MELODIC. This reduced the drug and placebo effect parameter maps to a set of 110 features (29, 36). For all assessments, we used a basis set generated from all studies except the study under assessment, to avoid any chance of bias.

The forced-choice task of identifying which of two sessions involves an analgesic is a ranking problem, which can be performed with a ranking SVM (26). A ranking SVM differs from an SVM classifier in that it aims to rank a set of samples in some way, rather than classify individual samples. The present task requires ranking the scans by identifying which of the two is most likely to involve the drug (for example, a forced choice between placebo and drug conditions). SVMs have been shown to deal well with large feature sets, and are consistently one of the best performing methods in imaging comparison studies (28). The C parameter of the SVM determines the trade-off between the width of the decision-boundary margin and the number of support vectors (or mis-classified points). We used a nested cross-validation procedure to identify an optimal C value. The Python Multivariate Pattern Analysis (PyMVPA) and Scikit-learn toolkits were used for prediction, with SVM routines from the LibSVM library (37–39). The forced choice classification meant that sensitivity and specificity are always equal to the prediction accuracy.

Assessing normalization of responses to clinical pain states

This assessment determined whether drug-induced modulations of stimulus responses resembled the differences in responses seen when the severity of the disease-relevant stimuli was reduced (e.g., the difference between different intensities of painful stimuli). For example, a

number of the test studies included painful stimuli to allodynic and normal skin, with the drug expected to modulate allodynia. A forced-choice SVM classifier was trained to distinguish between responses to two levels of disease-relevant stimuli under placebo conditions (for example, painful stimuli applied to sensitized and non-sensitized skin). This classifier was then tested on a held-out subject, with the responses associated with the lower symptom severity being replaced by responses to the higher symptom severity during the drug sessions (fig. S1). Positive evidence for efficacy was flagged if the drug-modulated responses were consistently identified by the classifier as responses associated with the reduced clinical symptoms.

Supplementary Figures

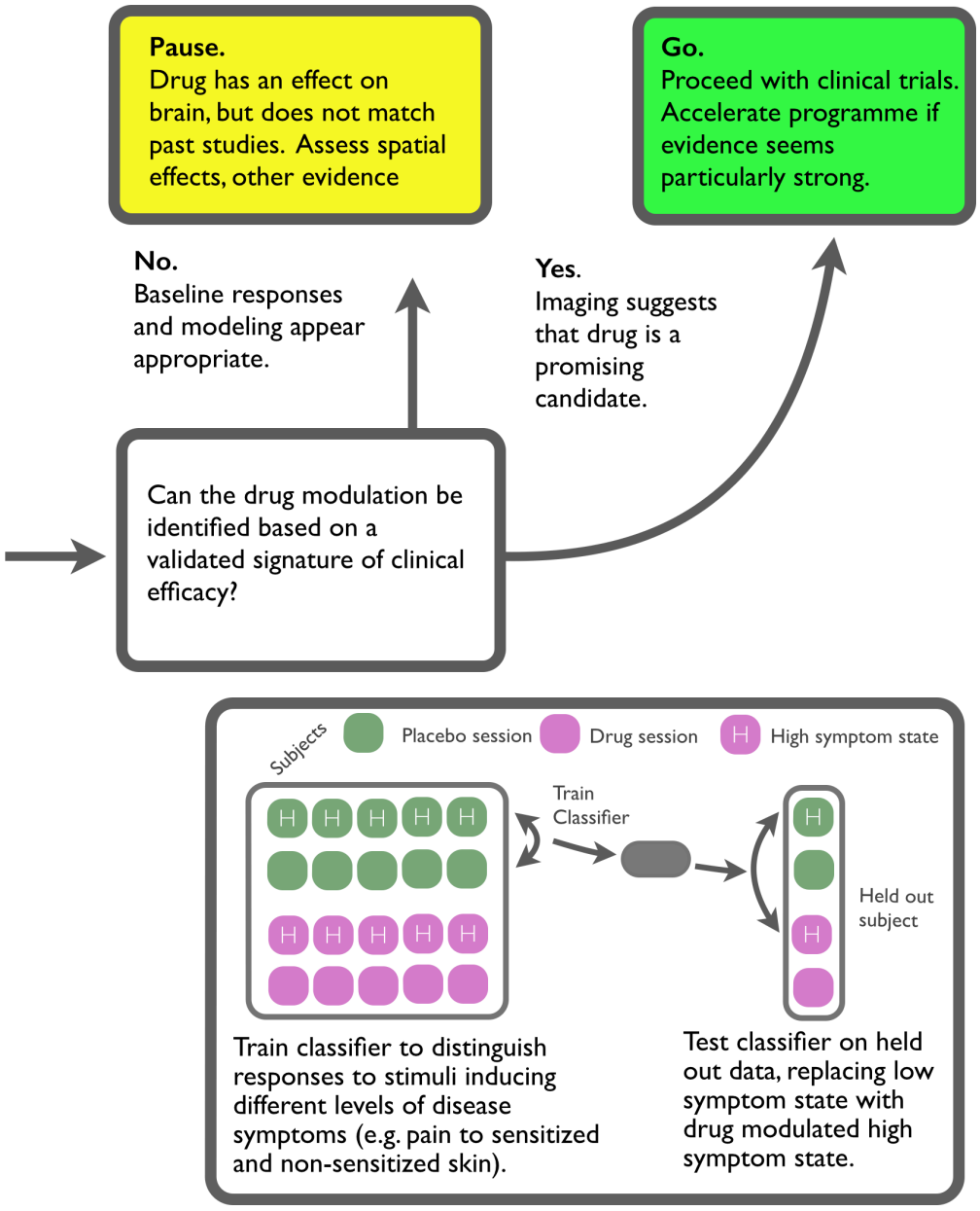

Fig. S1. Alternative clinical efficacy procedure. In this assessment, a classifier was trained to distinguish different levels of disease symptoms (for example, strong and weak painful stimuli). The classifier was then tested for its ability to distinguish drug sessions from placebo session responses to painful stimuli. If the drug session was consistently identified as the lower weaker stimulus, this was taken as evidence of efficacious action.

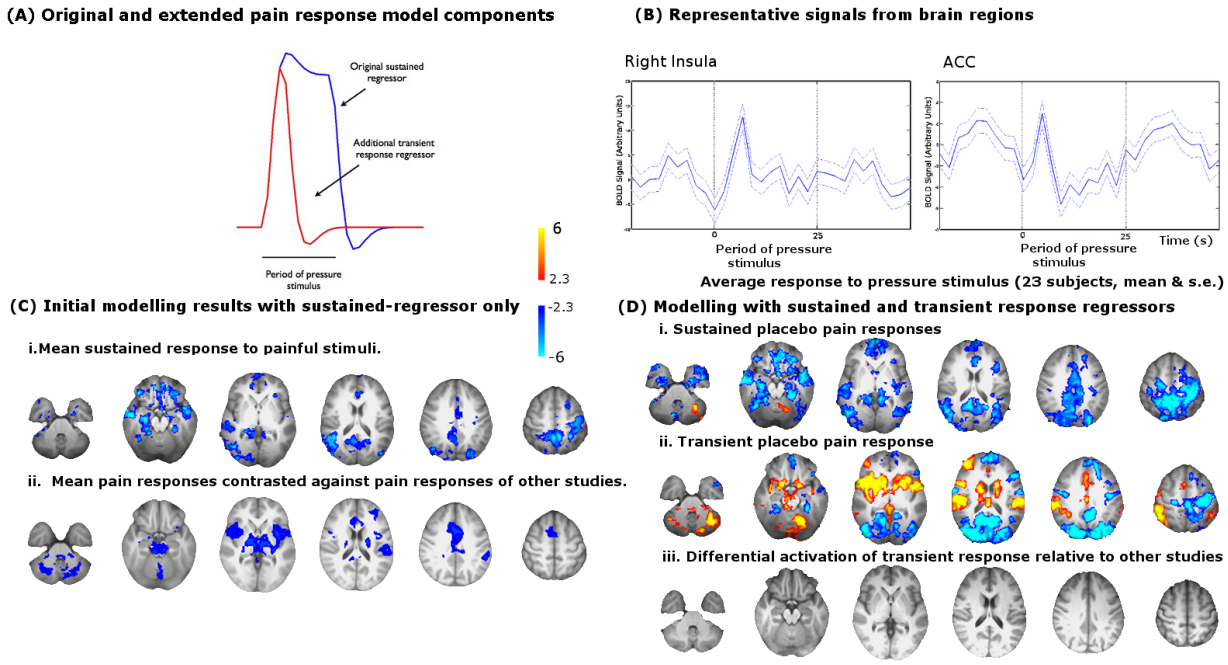

Fig. S2. Modeling of pain responses in the pregabalin study (b) (*27*). This study was flagged in QA as showing anomalous pain responses. Further assessment found the responses did not match the initial response model. **(A)** Pain response modeling. Blue line shows the original sustained response component. Red line shows the transient response component added after QA. **(B)** Time course of fMRI BOLD signal responses extracted from the imaging recordings the insula and anterior cingulate cortex (ACC) brain regions (n = 23 subjects; confidence intervals indicate standard errors). **(C)** Sustained component–only modeling of the pain response. **i.** Regions showing a significant sustained signal response for the full duration of the mechanical stimulus. All effects are deactivations of the BOLD signal relative to baseline. **ii.** Regions showing significantly lower pain responses compared to those seen across all other studies in the multi-study dataset (two-group t-test, cluster-based thresholding, blue corresponds to regions where study (b) was less than the multi-study average). **(D)** Significant effects identified using the extended pain response model (paired t-tests, where red/yellow is a positive response; blue is deactivation). No differences were found compared to other studies.

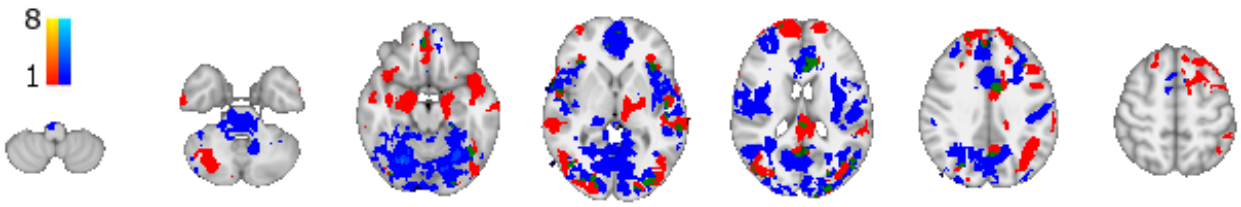

Fig. S3. Summary of significant effects in individual study placebo vs. drug pain response contrasts for individual analgesic studies in Table 2. Blue regions indicate where one or more studies had significant decreases in responses during drug conditions. Red indicates where one or more studies had greater responses in drug conditions. Green areas indicate where some studies showed significant increases and others significant decreases. Color bars indicate the number of studies showing significant effects. All contrasts were paired t-tests and statistically thresholded identically with cluster corrected $P \leq 0.05$.

SUPPLEMENTARY TABLES

Table S1. Results of procedure applied to analgesic studies when experimental stimuli with lower clinical relevance are used. Drug assessment protocol results when stimuli with lower clinical relevance than the primary stimuli (Table 2) were used as inputs to the assessments. Colors indicate whether compounds passed a specific assessment phase based on the predefined decision rules described in Results: Green –pass; Yellow – qualified pass (Q); Red – fail. Non-hyperalgesic: stimulus was to hyperalgesic skin. Non-allodynic: stimulus was to non-allodynic skin. *P*-values indicate the probability of achieving this accuracy or better given no drug effect (chance = 50%). CIs are 90% Wilson-score confidence intervals. Analgesic modulations were not as consistently identified as they were with the primary stimuli, with four studies failing the pharmacodynamic effect assessment.

Study	**Drug (condition)**	**QA** Area of reduced response (%)	**Pharmacodynamic effect** Accuracy (range)	*P*	**Clinical efficacy** Accuracy (range)	*P*	**Decision**
Additional study contrasts							
a	Gabapentin (non-hyperalgesic)	0	58% (36-78)	0.193	75% (51-90)	0.019	Go(Q)
c	Pregabalin (non-allodynic)	0	63% (43-79)	0.105	63% (43-79)	0.105	Go(Q)
c	Pregabalin (thermal)	0	38% (21-58)	0.773	63% (42-79)	0.105	Go(Q)
d	Tramadol (non-allodynic)	0	25% (12-45)	0.96159	68% (48-84)	0.038	Go(Q)
d	Tramadol (thermal)	0	38% (21-58)	0.773	50% (31-69)	0.402	Stop
e	Remifentanil (thermal)	0	82% (65-92)	0	79% (61-90)	0.002	Go
g	THC (non-hyperalgesic)	0	71% (49-86)	0.029	57% (36-86)	0.212	Go(Q)
f	Naproxen (mild pain)	0	74% (49-86)	0.010	79% (61-90)	0.002	Go

Table S2. **Outcome of the analgesic assessment protocol when using inputs derived from a set of pain-related ROIs.** Colors indicate whether compounds passed a specific assessment phase based on the predefined decision rules in Fig. 1: Green –pass; Yellow – qualified pass (Q); Red –fail. Grey – reassess data/analysis. Three analgesic studies were not successfully identified by this assessment protocol. QA characterized the percentage of voxels showing responses that were significantly lower than responses in regions consistently responding to similar stimuli in the other studies (voxelwise unpaired t-test, corrected for multiple comparisons). Accuracies give the proportion of subjects for which the analgesic session was correctly identified. *P*-values indicate the probability of achieving this accuracy or better given no drug effect (binomial test, chance = 50%). Sample sizes correspond to number of subjects reported in Table 1. 90% Wilson-score confidence intervals are shown.

Study	Drug	QA	Pharmacodynamic effect		Clinical efficacy		Decision
		Area of reduced response (%)	Accuracy (range?)	*P*	Accuracy (range?)	*P*	
			Analgesic study contrasts				
a	Gabapentin	0	75% (51-90)	0.019	75% (51-90)	0.019	Go
b	Pregabalin (I)	0	52% (36-68)	0.339	48% (32-64)	0.500	Stop
c	Pregabalin (II)	0	44% (26-64)	0.598	38% (21-58)	0.772	Stop
d	Tramadol	0	38% (21-58)	0.772	38% (21-58)	0.772	Stop
e	Remifentanil (I)	0	91% (76-97)	0.000	95% (82-99)	0.000	Go
f	Remifentanil (II)	0	92% (69-99)	0.000	92% (69-99)	0.000	Go
g	THC	0	71% (49-86)	0.029	43% (24-64)	0.029	Go(Q)
h	Naproxen	0	63% (44-79)	0.084	63% (44-79)	0.084	Go(Q)
			Control study contrasts				
i	Placebo - Study (b)	0	30% (22-54)	0.953	43% (32-64)	0.661	Stop
j	Placebo - Study (e)	0	59% (42-74)	0.143	59% (42-74)	0.143	Go(Q)
k	Baseline - Study (f)	0	67% (43-84)	0.072	42% (22-64)	0.612	Stop
l	Visual stimulus - Study (f)	79	91% (70-99)	0.000	8% (0-30)	0.997	Reassess
m	Auditory stimulus - Study (f)	6	50% (29-71)	0.387	42% (22-64)	0.612	Reassess
n	Visual stimulus - Study (h)	48	32% (17-50)	0.916	58% (42-74)	0.143	Reassess