Accepted Manuscript

Title: Pain Neuroimaging in Humans: a Primer for Beginners and Non-Imagers

Author: Massieh Moayedi, Tim V. Salomons, Lauren Y. Atlas

PII: S1526-5900(18)30122-6
DOI: https://doi.org/10.1016/j.jpain.2018.03.011
Reference: YJPAI 3559

To appear in: The Journal of Pain


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 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.
Pain neuroimaging in humans: A primer for beginners and non-imagers

Massieh Moayedi¹,²,³, Tim V Salomons⁴,⁵, Lauren Y Atlas⁶,⁷

¹Faculty of Dentistry, and ²University of Toronto Centre for the Study of Pain, University of Toronto, ³Department of Dentistry, Mount Sinai Hospital, Toronto, ON Canada; ⁴School of Psychology and Clinical Language Science and ⁵Centre for Integrated Neuroscience and Neurodynamics, University of Reading, Reading, UK; ⁶National Center for Complementary and Integrative Health and ⁷National Institute on Drug Abuse, National Institutes of Health, Bethesda, MD, USA

Short Title: A primer for human pain neuroimaging

Text Pages: 59

Number of figures: 3

Correspondence to:
Massieh Moayedi
Faculty of Dentistry
University of Toronto
124 Edward Street, Room 260 B
Toronto, ON
Canada M5G 1G6
P: 416-979-4900 x4364
E: m.moayedi@dentistry.utoronto.ca
https://www.dentistry.utoronto.ca/faculty-profiles/massieh-moayedi

DECLARATIONS:
LYA is supported by the Intramural Research program of the NIH's National Center for Complementary and Integrative Health and National Institute on Drug Abuse. MM is supported by an IASP Early Career Award, and the Bertha Rosenstadt Endowment. TVS was supported by a Marie Curie International Incoming Fellowship from the European Commission, and a British Academy Leverhulme Small Research Grant. The authors have no conflicts to report.
**KEYWORDS:** Pain, MRI, sMRI, fMRI, MVPA, functional connectivity, guidelines

**SUMMARY**

We provide a guideline for designing, analyzing, reading, evaluating, and reporting results of a pain neuroimaging experiment.
ABSTRACT

The field of human pain neuroimaging has exploded in the last two decades. During this time, the broader neuroimaging community has continued to investigate and refine methods. Another key to progress is exchange with clinicians and pain scientists working with other model systems and approaches. These collaborative efforts require that non-imagers be able to evaluate and assess the evidence provided in these papers. Likewise, new trainees must design rigorous and reliable pain imaging experiments. Here, we provide a guideline for designing, reading, evaluating, analyzing, and reporting results of a pain neuroimaging experiment, with a focus on functional and structural MRI. We focus in particular on considerations that are unique to neuroimaging studies of pain in humans, including study design and analysis, inferences that can be drawn from these studies, and the strengths and limitations of the approach.

Perspective: This article provides an overview of the concepts and considerations of structural and functional MRI neuroimaging studies. The primer is written for those who are not familiar with brain imaging. We review key concepts related to recruitment and study sample, experimental design, data analysis and data interpretation.
Table of Contents

1. Introduction ...................................................................................................................... 4

2. Recruitment and sample .................................................................................................. 6
   2.1. Statistical power and sample size ........................................................................... 9

3. Experimental design ...................................................................................................... 11
   3.1. Consider the context: constraints of the MRI environment ..................................... 11
   3.2. Materials and procedures ...................................................................................... 13

4. Analysis .......................................................................................................................... 16
   4.1. Preprocessing ....................................................................................................... 16
   4.2. Subject-level task-based fMRI statistical analyses ................................................. 18
   4.3. Subject-level resting state fMRI statistical analyses ............................................... 21
   4.4. Group-level statistical analysis .............................................................................. 23
   4.5. Multivoxel-pattern analysis .................................................................................... 27
   4.6. Functional connectivity analysis ............................................................................ 27
   4.7. Model-based analysis ............................................................................................ 31

5. Interpreting Data and Making Inferences ........................................................................ 32
   5.1. Inference ............................................................................................................... 32
   5.2. Logical fallacies in neuroimaging ........................................................................... 34

6. Conclusion ..................................................................................................................... 36

7. References ..................................................................................................................... 37
1. Introduction

Understanding how pain is encoded in the brain has been a fundamental challenge for pain researchers. Despite the universality of acute pain and the high prevalence of chronic pain, we have yet to precisely characterize the mechanisms of pain perception and modulation in health and disease. The complexity of identifying these mechanisms stems from the multidimensional nature of pain – pain is a complex amalgam of sensory, affective, cognitive, and motor responses to dynamic internal and external states. The challenge of characterizing mechanisms increases as pain becomes chronic, with widespread plasticity of nociceptive and modulatory pathways contributing to the ongoing experience of pain. The rise of neuroimaging techniques offers the potential for breakthroughs in these efforts.

Neuroimaging approaches (including functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), and other approaches) are extremely powerful tools that offer unique, non-invasive, in vivo views of central processes. Indeed, functional and structural neuroimaging studies have identified neural responses and features of both acute [4,46,57,68,76,180,225,226], and chronic pain [3,33,57], as well as effects of both pharmacological [9,29,30,69,198,230] and non-pharmacological [27,28,48,72,126,131,171,203,217,223,226,227,239] interventions.

Due to the promise of these new techniques, pain imaging is a rapidly growing field that will continue to expand as scanning facilities become more available and analysis software becomes increasingly user-friendly. While brain imaging findings can provide important insight into central mechanisms, there are many aspects of the study design and analysis that must be carefully considered and planned a priori to obtain a robust, reproducible result. Indeed, a recent systematic review of fMRI data found that a dataset could be processed through almost 7000 unique pipelines, with almost 35,000 resulting maps [38]. This highlights the importance for non-imagers and new trainees who read neuroimaging papers to be familiar with some of these considerations and how they may affect outcomes and inferences. Here, we will provide an overview of these considerations, and key questions that should be asked when reading an imaging paper. We will not weigh
in on the many exciting theoretical debates in the pain neuroimaging community, such as the specificity of pain-evoked cortical and subcortical responses or the feasibility of brain-based biomarkers for pain (see, for example: [56]). We instead provide a primer and guidelines to assist trainees and non-experts in designing and reporting pain neuroimaging experiments as well as reading and evaluating manuscripts.

Given the brain’s key role in generating pain percepts, the ability to noninvasively examine brain function in vivo is critical. We focus specifically on functional and structural MRI, as MRI is the most common tool used in human pain studies. Other promising human neuroimaging methods such as electroencephalography, magnetoencephalography, functional near-infrared spectroscopy, transcranial magnetic stimulation, positron-emission tomography, and arterial spin labelling are beyond the scope of this review.

A cursory PubMed search with the search terms “(pain or nocicept*) AND brain AND mri NOT review” on January 20, 2017 resulted in 4895 studies (see Figure 1). A number of publications address methodological issues and statistical considerations associated with human neuroimaging in general [39,145,147,148,169,173,181,185], and we encourage neuroimagers to consult these reviews for additional guidelines, and more in-depth discussion of some of the technical considerations we delineate below. Our aim is to address how these issues may specifically impact pain research. Our goal is to provide an introduction of particular use to a novice audience, including new trainees, clinicians, and/or non-imagers interested in evaluating studies on the neuroimaging of pain. We believe that awareness of methodological and inferential limitations can lead to positive advances in the field. We focus on the elements that should be included in any paper’s Methods and Results sections, and address inferences that may be drawn during the Discussion.

2. Recruitment and Sample.

Sample characteristics. After deciding on a research question, the first aspect any researcher considers is his or her research population and sample. It is important to acknowledge that neuroimaging experiments require unique constraints on enrollment, and
many of these constraints might put a particular burden on neuroimaging studies of patient samples.

MRI scanners are large magnets with a narrow bore in which the participant lies during scanning. The participant’s brain is scanned with specialized coils embedded in a small headcage, or headcoil. This provides several constraints that are rarely acknowledged but can substantially impact a study. Subjects must fit in the narrow bore, and therefore MRI studies are likely to exclude obese patients, despite the fact that obese individuals are more likely to report severe pain than normal and under-weight counterparts [109]. Because the MRI scanner uses radiowaves, participants cannot have any ferromagnetic metal in their body, as there is a risk of the main magnetic field pulling on the metal, especially as the subject enters the bore of the scanner. Radiowaves can also cause heating in tissues, and this can be exacerbated by any electric conducting materials including cables and wires. Furthermore, metal in the head and neck region can cause image quality issues, as these reflect the radiowaves. These factors provide severe constraints on researchers interested in studying topics such as phantom pain in amputees or pain in individuals after certain surgical procedures, as these patients may have metallic (and possibly ferromagnetic) objects in their body that are not safe in a research MRI scanner. Studies of pain in other populations, such as the elderly, or patients with cardiac pain, may also be limited due to other types of incompatible implants or devices, including pacemakers or medication pumps. Finally, some patients may experience claustrophobia because of the narrow bore or confining coil surrounding their head, and so patients with comorbid anxiety disorders may be less likely to complete these experiments. These constraints mean that such studies might routinely exclude the most severe cases, which must be considered when drawing general inferences about pain populations. Studies must always measure and report complete characteristics of the sample as well as inclusion and exclusion criteria.

Given the constraints on the participants that can be recruited for an imaging study, it is often tempting to recruit a convenience sample – usually healthy young adults from the university environment. However, such a sample is not representative of those
disproportionately affected by chronic pain: women between the ages of 30 and 50, of lower 
socio-economic position [206]. Nonetheless, many neuroimaging experiments are not 
designed to isolate mechanisms of chronic pain, and may instead be interested in more 
general neural mechanisms, such as those involved in the psychological modulation of pain, 
or basic processes of nociception and acute pain. Convenience samples may be appropriate 
in these cases, as they allow investigators to isolate basic mechanisms underlying pain 
perception in healthy individuals. However, even studies of acute pain in healthy volunteers 
should carefully consider generalizability to the larger population, as convenience samples 
may lack diversity in age, educational background, socioeconomic status, race and ethnicity, 
depending on the study, and each of these factors has been shown to influence pain 
[36,66,125]. All experiments should report full sample characteristics so that readers and 
reviewers can evaluate generalizability to the population.

Many experiments compare patients with pain disorders to matched control patients 
and compare results across groups. If healthy controls are not selected to match non-pain 
related characteristics, there is a strong chance that differential findings between groups will 
reflect processes other than pain disorder of interest (e.g. group differences due to higher 
incidence of mood disorders, obesity and comorbid chronic pains) [216,219]. It is strongly 
recommended that such matching be done by carefully selecting the control population, 
rather than attempting to find “clean” pain patients (e.g. those free of co-morbid 
psychopathology). While capturing processes related to comorbid disorders complicates 
interpretation, and can lead to confounds, individuals without such comorbidities may not be 
representative of most pain patients. Potential confounds and comorbidities should be 
carefully considered, and if they are unavoidable, accounted for with appropriate 
experimental designs and statistics. Researchers must also carefully consider differential 
artifacts related but not germane to clinical presentation. For example, recent work indicates 
that resting state fMRI (rs-fMRI) studies are highly susceptible to even small motion artifacts, 
such that a group difference could emerge if there was a systematic difference between the 
groups in movement [188,204] (as discussed in section 4.1). Factors such as discomfort
from lying in the scanner, comorbid movement disorders, or psychiatric disorders such as anxiety might be more likely in pain populations and lead to greater motion in patient groups, resulting in differential between group effects related to movement rather than the measure of interest [204]. Thus, head motion and potential group differences in motion must be carefully considered and quantified in both rs-fMRI and task-based fMRI studies, and data cleaning strategies should be carefully used to mitigate the contributions of such factors (e.g., [204]).

2.1. Statistical power and sample size.

All studies must ensure adequate sample size and statistical power to reliably detect meaningful effects, and neuroimaging experiments are not unique in this regard. Neuroimaging studies are expensive to run, generally requiring hundreds or even thousands of dollars in scanning fees for each subject. It is therefore common for sample sizes to be smaller in neuroimaging studies than other types of experiments. While small samples are common due to the financial burdens and additional constraints on neuroimaging studies, this has led to the publication of many studies that are likely to be underpowered for effects that might be of greatest clinical interest, such as individual differences in pain populations. Underpowered studies are less likely to detect true effects and more likely to find false positives by overfitting data, thereby decreasing the likelihood that findings will replicate [34]. Studies of chronic pain populations might be particularly susceptible to such effects given the breadth of diagnostic categories (meaning that patients with similar diagnoses but divergent symptom profiles might be included in the same small sample; e.g., fibromyalgia which is commonly a diagnosis of exclusion [79], where symptom profiles can vary greatly) and heterogeneity in terms of comorbid disorders and medication use.

Given the potential for false positives in underpowered imaging studies, studies with small samples are coming under increased scrutiny. The onus will increasingly be placed on researchers to demonstrate that samples are adequately powered to detect expected
effects. Such demonstrations must be based on justifiable, *a priori* calculations, so researchers should always estimate power prior to conducting an fMRI study, and report how the sample size was determined. Statisticians have recently introduced several new, easily accessible fMRI-specific power analyses that make it easier to consider desired effect and sample sizes (e.g. fmriPower.org [168], neuropowertools.org [71]) and will help to justify the funding of fully powered studies. These approaches can be used *a priori* to estimate sample size based on expected effect size. Importantly, a number of these approaches require researchers to have pre-existing imaging data, which is not always possible for new researchers or those using new paradigms. Power calculation based on behavioral effect sizes may also be useful as these do not rely on previous imaging studies. These can be computed using data from pilot studies completed outside of the fMRI environment, which would substantially reduce costs.

The problem of underpowered studies can also be addressed by aggregating data via repositories and/or meta-analyses. Repositories facilitate data sharing across imaging experiments. Some repositories host data from many different tasks, scanners, and populations. These repositories facilitate reproducibility, open science, and large-scale analyses that are relatively impervious to noise created by inconsistencies across individual experiments. Other repositories require contributors to collect brain images with standardized scanning protocols. These scans then undergo standardized preprocessing and analysis pipelines, allowing several groups with limited resources to pool brain imaging data and investigate larger cohorts of patients. Several pain imaging repositories (OpenPain – http://openpain.org; Pain and Interoceptive Imaging Network PAIN– https://www.painrepository.org [137]) have been established, and have proven to be successful. However, there are legal and ethical considerations that may affect a group’s ability to contribute to such repositories.

Finally, meta-analyses statistically test the distribution of findings across studies, which permits assessments of the consistency of effects and overcomes some of the
limitations associated with small individual studies. These studies allow for a principled approach to determining the relationship between a particular brain region and behaviour. Several pain imaging meta-analyses have been published, including meta-analyses of pain-evoked cortical responses [68,180], placebo analgesia [1,8], automated meta-analyses of brain activation associated with the term “pain” [235], and meta-analyses of brain responses and structural brain abnormalities in patients with chronic pain [40,80,144].

3. Experimental design.

3.1 Consider the context: constraints of the MRI environment.

Neuroimaging experiments (especially MRI) take place in a unique environment, so studies must be carefully designed and made suitable for the imaging suite. As mentioned above, this restrains the patient populations that can be studied. It also places substantial constraints on the types of experiments pain imagers can conduct. All equipment must be MRI-compatible and suited to the unique scanner environment. There are few commercially available MRI-compatible devices capable of delivering nociceptive stimuli, as these must be completely non-ferromagnetic and must not introduce electrical noise during data collection. Even commercially available devices may have differential success at different scanners depending on considerations such as MRI field strength, sequence design, and even bore size. Thus, many pain researchers have examined brain responses associated with acute thermal and electrical pain, but few have examined cold pain, cold allodynia, or chemical pain. During the experiment, painful stimuli are usually controlled by simple computer tasks that coordinate stimulus presentation and synchronize timing with the MRI scan (see Experimental Procedures, below). Experimenters often use these programs to present visual stimuli (e.g. task instructions, cues) and to measure responses (e.g. pain ratings). Visual stimuli are usually presented on a computer screen that patients view through goggles or via a mirror that sits atop the headcoil and reflects images displayed on a projector screen. Participants provide pain ratings and make other responses using devices such as button
boxes, joysticks, mouse-like trackballs, or by moving their hands [5] since verbal responses are generally avoided due to the loud noises produced by the scanner and because head motion must be minimized. Thus, fMRI studies of pain may not capture the social and interpersonal elements that are likely to contribute to verbal pain ratings when a patient informs her doctor about her pain. This can be seen as a strength (e.g. for researchers who investigate ascending nociceptive pathways and want to minimize social modulatory influences) [60,141,142,212] or a weakness (e.g., patient-provider interactions are thought to be a critical component of placebo analgesia [65,77,119,121], and few computer paradigms are able to capture these processes).

Head motion provides a third unique constraint afforded by the MRI scanner context. Participants cannot move their heads more than a few millimeters (typically less than 2mm) during scanning, for fear of contaminating the data (see below). Some individuals, such as those with lower back pain, may not be able to lie still without pain, which provides additional constraints on patient samples and eligibility. In addition, painful stimuli that induce strong withdrawal responses such as startle and electric shock cause unique challenges due to task-related motion if not carefully mitigated [37,238] (see 4.1 for discussion of motion correction). Many researchers choose to familiarize subjects with the stimuli to minimize startle and other withdrawal-related behaviors, as well as to ensure stimuli are tolerable for the participants. This of course imposes limitations on the intensity of pain or novelty of the stimuli that can actually be administered in the scanner. Furthermore, it is unclear to what degree neural responses might reflect regulation of the prepotent withdrawal response even when the participant is able to voluntarily suppress motion. All of these considerations must be weighed in terms of the construct validity of painful stimuli presented within the scanning environment.

Another key consideration is the scanning parameters and sequences. Most sites will have "out of the box" sequences, but these should be selected with consideration, in order to best optimize the sequence to maximize signal from regions of interest. Although the technical aspects of these decisions are outside the scope of this review, it should be
recognized that these considerations must be tailored to experimental hypotheses. We therefore recommend that researchers work closely with MR physicists and experts to ensure that the selected sequences are ideally suited to the temporal resolution of their effect of interest and the particular anatomical region(s) of interest. For example, some brain stem structures might require multiple acquisitions of different contrasts for precise anatomical localization. Additionally, protocols might be adjusted for experimental efficiency. For example, a standard diffusion-weighted study to investigate white matter in the brain might acquire 64 diffusion-encoding directions, with b-shell of 1000 s/mm$^2$. However, this acquisition is more than might be needed for a standard FA map, which requires a minimum of six diffusion-encoding directions [17,18,166], resulting in an unnecessary cost and participant burden. These study specific considerations might also be balanced against the homogeneity requirements of data sharing repositories, although collaborative and multisite studies also exist (e.g., the Multi-Disciplinary Approach to the Study of Chronic Pelvic Pain; [45,139]) which may permit more heterogeneity.

3.2 Materials and procedures
Authors should outline everything that participants did from the start of the study until they finished. A study begins once a participant provides informed consent, as approved by a local Institutional Review Board, which evaluates the ethics of the study. Informed consent and ethics approval should be acknowledged in any study, and authors should include important details, for example whether authorized deception was used in any studies that include misleading information (e.g. studies of placebo analgesia), or whether patients were asked to refrain from taking their prescribed medication. In some cases, understanding the relative rate of participation is important and studies should report the number of people contacted for recruitment. Were subjects debriefed at the end of the study? For studies of individual differences and/or clinical severity, which questionnaires were administered (not just the ones relevant to the current paper)? Similarly, were any tasks or relevant procedures administered outside of the main experiment that are not analyzed in the paper? This is
important information as additional tasks and measures might influence behavior in the main paradigm, and reviewers and readers should be able to evaluate potential confounds.

The description of task design should include all details necessary for the purpose of external replication, including the instructions, counterbalancing schemes, and task timing. What platform was used for experimental programming, and how did subjects provide responses? If decisions about task design were made based on earlier pilot testing, it can be useful to report these details (e.g. “We collected pain ratings 20 seconds after heat offset. Pilot testing revealed that there was no difference between ratings made immediately after offset versus after a delay. We chose the current design so as to reduce contamination of the BOLD response to heat”). It’s also important to provide the exact instructions provided, as well as a description of the scales used to assess painful percepts, and other features of the stimulus. What were the exact instructions? What was the resolution of the scale? What were the anchors to the scale?

Some pain studies use nociceptive stimuli to elicit pain in healthy subjects and/or in chronic pain patients. Researchers can choose between several nociceptive stimulation paradigms. Thermal pain is most well-established and most common stimulus used in neuroimaging [68]. This is likely because the paradigm works well in the scanner and is relatively convenient. However, it is a poor model of chronic pain, as heat pain is qualitatively dissimilar to pain experienced in most chronic pain conditions. Therefore, the researcher must ensure that inferences are appropriate. It should not be taken for granted, for instance, that acute heat pain is a valid model for chronic pain that is neuropathic or musculoskeletal in nature. Other acute pain models such as ischemia, muscular hypertonic saline injections, bladder filling and visceral/rectal distention might be more appropriate for questions about the neural bases of musculoskeletal or visceral chronic pain disorders [6,55,96,132,134,178,179,215]. However, not all pain experiments seek to model chronic pain; some focus on understanding neural processes associated with acute pain and its modulation. In this case, different considerations guide the evaluation of validity. Are experimental manipulations appropriate with regard to the psychological construct the
researchers are testing? Is the chosen noxious stimulus appropriate for the questions of interest (e.g. thermal vs. electrical vs. laser pain activate different ascending fibers)?

Questionnaires are also often an important component to a study. These can serve as screening tools against exclusion criteria (e.g., handedness, other neurological disorders, and MRI contraindications, such as claustrophobia), to characterize a chronic pain population (e.g., the McGill Pain Questionnaire [159,160]), or as a measure or covariate of interest. When used judiciously, such measures can help strengthen inferences about the mental processes that particular patterns of activation might be subserving. On the other hand, each added questionnaire inflates the number of statistical tests and increases the chance for false positives, if not carefully taken into account through multiple comparisons correction. As discussed below (see 4.4), the issue of type-1 error is particularly germane to neuroimaging studies (see below), and increases patient burden in terms of time spent on the study. Thus, the type and number of questionnaires included should be carefully considered on the basis of clear hypotheses. Questionnaires expected to have large amounts of overlapping variance (e.g., the Fear of Pain Questionnaire [158] and the Pain Catastrophizing Scale [211]) can increase the chance of incidental findings with little additional explanatory power added to the study.

There are three different approaches that can be used to investigate neuroimaging data: hypothesis-driven, exploratory, and reproduction/replication. A typical neuroimaging study uses these respective approaches to test a hypothesis, develop new hypotheses, and confirm (or reproduce) the findings. Hypotheses should be selected \textit{a priori}, before any data is collected, on the basis of theory and prior data. Once a hypothesis is generated, a pilot study is usually performed on a small sample size to ensure that the experimental paradigm is valid (i.e. to ensure that the hypothesis is appropriately operationalized), and to determine the feasibility of the study. Small details may be adjusted at this time (e.g. instructions, task timing, etc). Pilot studies are rarely submitted for publication, as they often use liberal statistical thresholds (due to low power). They allow researchers to adjust the paradigm prior to fMRI scanning, as scan time is costly and one wants to make sure the task is clear so that
scanning can proceed uninterrupted. Pilot studies can also be used to estimate effect size, although it is known that small samples can overestimate effect size [34]. Then, an independent sample (whose size has been determined using a power calculation) consisting entirely of new subjects should be acquired to formally test the hypothesis. Preregistration allows researchers and reviewers to distinguish between a priori, hypothesis-driven analyses and post hoc exploratory analyses, which are intended to generate new hypotheses. Several preregistration sites now exist (e.g. Open Science Framework https://osf.io/registries/; [177]), and several neuroscience and psychology journals now allow researchers to conduct preregistered reports (https://cos.io/rr/). We hope that interest in preregistration will grow within the pain community. We recognize that due to an emphasis on novelty among publication and funding outlets, the choice to follow a path from exploratory to replication studies is generally not in the hands of young or even senior investigators. However, given the paucity of replication studies in fMRI, and its reliability being called into question [23,54], we join others in advocating for a funding and publication climate where confirmatory neuroimaging studies are encouraged [11,165].

4. Analysis

Here we provide an overview of the analyses that should appear in a typical imaging paper and provide some guidance about how choices might affect pain studies. We focus primarily on the Results section of a typical fMRI-BOLD investigation, as these are the most prevalent type of studies, although we acknowledge structural MRI studies where relevant. Other modalities may have different analysis steps, and specific literature should be consulted. Furthermore, trainees should refer to more comprehensive papers on how to analyze and report neuroimaging studies [174,185].

4.1 Preprocessing

Once data are collected, several steps are necessary to transform the data into the proper multi-dimensional format that can undergo statistical analysis. These transformations
are grouped together and generally referred to as preprocessing. The fMRI pipeline usually includes steps to (a) remove the first set of volumes acquired (usually between 5-10), as it takes some time for the magnetic field to reach a steady state; (b) correct for the fact that fMRI data are collected in slices, and slice time acquisition differs within each time point ("slice time correction"); (c) test whether the head moved at any time point, which can induce artifacts, and correct for any head motion so that analyses capture to the same brain region over time ("motion correction"; see [94,95,106,186,237] for in-depth discussion of the problem with motion in MRI, and reviews of approaches to motion correction); (d) remove high- or low-frequency noise that can contaminate the signal or lead to spurious results ("temporal filtering"); (e) register the subject’s functional images with structural images (since the images are collected separately in time and head displacement might occur; "coregistration"); and (f) spatially transform the individual’s images to a standard template brain (e.g. MNI template, Talaraich-Tournoux atlas, a group mean) with a specific stereotaxic space so that analyses are possible across individuals, who vary in neuroanatomy ("normalization"). The normalization step is also important for spatial specificity when reporting results of an fMRI study, and allows findings to be interpreted by other researchers and be included in subsequent meta-analyses. Several additional optional steps might occur at this stage (e.g. correction for artifacts, spatial smoothing, etc.), but details of these individual procedures are outside the scope of the current review [82,112,186].

Preprocessing sMRI data for gray matter analysis largely comprises of (1) removing scanner-induced noise from the structural T1-weighted brain images, and (2) segmenting tissues based on image contrast, and/or model the tissue of interest. Finally, the images are transformed and registered to a stereotaxic coordinate space for analysis. For example, for a gray matter analysis the B1 field is calculated to remove signal inhomogeneities. This allows for better tissue classification, and better estimates of gray matter volume or cortical thickness. For diffusion-weighted scans, which are used for white matter analysis, susceptibility artifacts, such as eddy currents induced by gradients coils, are corrected using
various algorithms. Next a diffusion model (whether the tensor model to calculate fractional anisotropy, or a tractographic model) is calculated and applied to the data.

It is important to note that many of the algorithms used for these specific steps may vary as a function of software package (i.e. SPM, vs AFNI vs FSL), as well as the fact that individual programs exist to optimize each of these steps. We highly recommend that new researchers/trainees carefully investigate the specifics and defaults of the analysis software. Default assumptions are not always transparent, and may not always suit your individual needs. For example, if a researcher is interested in effects in small regions like the periaqueductal gray or the hippocampus, he or she might want to minimally spatially smooth the data (which blurs the boundaries of image units (3D pixels, or “voxels”), rather than use the default values within a software package. Large smoothing kernels are suitable for analyses of cortical regions, but not for smaller, more discreet brain regions [194]. As another example, SPM’s default settings can restrict statistical analyses to regions where the calculated signal is greater than a set threshold. This can particularly impact voxels that are prone to signal loss from susceptibility artifacts in regions with tissue-air boundaries [43], such as the inferior temporal lobe and the orbitofrontal cortex and other ventral brain areas, which are of interest in many pain studies. These artifacts can be reduced through informed decisions about acquisition parameters [64,92,189,208,224]. Neuroimagers benefit from developing expertise across multiple analysis packages and making informed decisions about which approach to use for a given analysis. Imagers should also develop practices that involve visualizing data at each stage of analysis, which can help to identify such issues as substantial dropout, missing data, and poor normalization or coregistration, among other important steps. Of course, when researchers elect to apply specific approaches, they should report their decision process and the reason that they opted for a given technique.

The following sections describe analysis considerations unique to task-based fMRI (4.2) and rs-fMRI (4.3).

4.2 Subject-level task-based fMRI statistical analyses.
Once fMRI data have been preprocessed, they can be analyzed in relation to the tasks that were administered during the fMRI session. The simplest way to think of the most common fMRI subject-level analysis approach, referred to as the General Linear Model (GLM), is that we test the correlation between each voxel's activity and the events that occurred during the experiment. To accomplish this, the researcher must carefully track the timing of all the events in scanning session relative to the start of scanning, including the stimuli delivered (e.g. noxious stimuli, images) and any responses collected. It is also important to record responses required for the analysis (e.g. pain ratings). A model representing the onset and duration of the various stimuli, a timecourse of the scan, is constructed (Figure 2). This stimulus timecourse will then be transformed to represent an ideal response to a given stimulus by combining the moments the stimulus was presented with a wave-like function that captures the biological delay in the BOLD response, referred to as a hemodynamic response function, or HRF. There are several options for HRF models, and these require an a priori understanding of how hemodynamics vary based on the population studied, stimulus type and the brain region of interest. For example, HRF in the brainstem, an important region when studying descending modulation of pain, may differ from those in the cortex [99,127], and the HRF shape and timings are different between young adults and the elderly [53]. When the events are combined, or “convolved”, with the HRF, this generates an example timecourse of what the BOLD response would look like if activity within a voxel activation increased every time this stimulus was delivered (see Figure 2). Different software packages model the HRF differently, and a researcher should understand how these differences can affect their results.

The researcher can simultaneously model several different stimulus types or conditions at this stage, and can test whether the magnitude of the response varies as a function of another variable – e.g. whether responses to a noxious stimulus are larger on the trials when the participant rated the stimulus as more painful. This is called a “parametric analysis” [120]. The different independent variables, or regressors, included in the GLM depend critically on the a priori design of the task, and the comparisons that the researchers
planned to make. To enhance the ability to measure task-related activation (i.e. activation that correlates with the design-based regressor created through the convolution steps described above), researchers can include nuisance variables, or regressors of no interest (e.g. head motion, intercepts for each block of data acquisition in case mean activation varies when the scanner stops and starts), to capture the noise in the MRI signal. Once the design matrix is complete, it is regressed against every voxel in a “mass univariate” approach [88] for each of the many-hundred thousand or so voxels that make up standard fMRI whole brain acquisition. Effectively, this means that an analysis is performed for each voxel in the brain – leading to over 100,000 statistical tests, and thus a high potential for false positives (see section 4.4 for discussion on correction for multiple comparisons). This subject-level analysis results in a coefficient (‘beta”) for each voxel for each element of the design matrix (all stimulus conditions and all nuisance variables). The coefficient describes the strength of the relationship between that regressor (the experimental condition or conditions of interest) and the voxel’s timecourse of activation. This usually involves averaging across stimuli and tasks, although some pain researchers use single trial analyses [170,190] to generate one beta for each event (per voxel) [7,8]. Researchers can also compute contrasts across these regressors at the subject-level (e.g. to compare high and low intensity stimulation) to generate voxel-wise contrast values that describe whether a given voxel’s activation strongly differs based on condition. Whole brain summary statistic maps of these voxel-wise beta coefficients or contrast values are then passed to group-level analyses to facilitate statistical analyses across groups of subjects.

4.3 Subject-level resting state fMRI analyses.

Rs-fMRI is becoming increasingly popular in the pain neuroimaging community, as it allows the investigation of intrinsic brain functional connectivity associated with pain-related phenotypes, including both pain-related characteristics (e.g. severity, duration) and cognitions (e.g. pain hypervigilance, fear of pain, catastrophizing) [12-
The preprocessing steps for rs-fMRI are similar to those of task-based connectivity, with a few notable exceptions. Rs-fMRI does not rely on a task-based regressor. Rather, time courses from spontaneous activity serve as a regressor. There are two primary approaches to the resting state analysis: seed-based connectivity or independent components analysis (ICA). The former relies on an *a priori* interest on the connectivity of a particular brain region, whose timecourse becomes the regressor. This approach requires additional nuisance regressors to correct for physiological noise, such as cardiac pulsatility and respiration-related motion. There are several approaches to correct for such noise [35], such as (1) measuring heart rate and breathing rate and including these measures as regressors, (2) post-hoc correction for physiological noise (retroICOR) [93], (3) regressing out timecourses from white matter and CSF (i.e., aCompCor) [20], or (4) global signal regression. In contrast, the ICA method does not rely on the timecourse of a particular brain region, but rather is a data-driven approach that groups correlated regions. In this method, artifacts can be identified based on the spatiotemporal patterns of the resulting components, and artefactual components can be removed from the signal. Once summary statistics maps of connectivity have been produced, a group-level analysis can be performed, similar to task-based data.

### 4.4 Group-level analyses

Once each subject’s fMRI data has been preprocessed, and statistics are performed on individual datasets, a group-level analysis is undertaken. The purpose of this analysis is to identify brain regions that are significantly activated across participants or between groups. At this stage, univariate statistical tests can be performed, or multivariate statistics can be employed. We focus primarily on traditional univariate statistics, which are tests that identify voxels in which there is significant activation as a function of the condition (or contrast) of interest. Perhaps the most common group-level analysis is the one-sample t-test on contrast maps across all participants to test whether activation at each voxel differs significantly as a function of the contrast (for example, whether all participants show greater activation in
response to noxious, relative to innocuous, stimuli). This generates a whole-brain statistical
map (e.g. a distribution of t-statistics across voxels), which must then be appropriately
thresholded for inference as described below. Other standard group-level analyses include t-
tests, which compare activation across groups (e.g. patients vs. controls), and correlation
analyses, which identify the strength of the correlation between voxel-wise activation and
individual differences in some known parameter (e.g. behavioral performance, symptom
severity, or questionnaire measure).

As with any type of statistical analysis, one must ensure results are not driven by
outliers. This is particularly important in small samples, and brain-behavior correlations, but
is true of all experiments. As an example, with a small sample size, a correlation between
pain duration and the thickness of the cortex in a particular brain region might be driven by a
single subject, or a small subset of subjects. Results should be visualized – i.e., scatter plots
should be inspected to determine whether effects can be attributed to a small number of
individuals. Alternatively, automated statistical techniques such as robust regression
[2,21,89,222] can reduce the contribution of outliers.

The group-level analyses outlined above, when conducted on each voxel throughout
the entire brain, require tens of thousands of statistical tests, i.e. multiple comparisons.
Given the inherent nature of statistics, this will necessarily lead to a cumulative proportion of
false positives. Therefore, it is important to adjust or correct for this inflated rate of false
positives, i.e., correct for multiple comparisons. There are two fundamental approaches to
the multiple comparisons problems: (1) region of interest (ROI) analyses, which reduce the
number of multiple comparisons based on a priori hypotheses, or (2) multiple comparisons
corrections, which correct for the number of tests performed. These are not mutually
exclusive, as testing multiple ROIs will still require multiple comparisons correction. We also
note that thresholds can be computed at the level of the voxel or the level of the cluster (for
a review of thresholding methods, see: [186]). We describe each of these considerations
below.

To reduce the number of multiple comparisons, researchers can reduce the search
area within the brain and restrain analyses to *a priori* ROIs. This excludes statistical tests from brain regions we do not expect to be implicated in the analysis, and reduces the number of comparisons, as tests are restricted to voxels within ROIs. When an analysis is performed on every voxel within an ROI, multiple comparisons corrections must still be conducted, but there will be substantially fewer comparisons for which to correct.

Researchers can also extract various coefficients across the ROI, depending on their question of interest. For example, the mean of the timeseries across all voxel in the ROI can be extracted. Another, and perhaps better, option is to extract the first principle component of the timeseries in the voxels. The resulting eigenvalue is effectively a weighted average of the activation in the ROI, where atypical voxels are down-weighted. Once the metric of interest is extracted, a single statistical test can be conducted, which would therefore be evaluated with a standard p<0.05 statistical threshold if data are combined (e.g. averaged) across voxels within the ROI. However, if coefficients are extracted from multiple ROIs or multiple voxels within an ROI, multiple comparisons correction is required. So how do researchers identify appropriate ROIs? One way is to include a functional localizer – an fMRI scan or contrast that excludes the condition of interest. For example, if we want to determine how a painkiller affects pain-related brain activation, a baseline scan of pain-related activation, in which each participant receives high and low intensity stimulation, can be performed prior to analgesic administration. This generates a functional localizer scan, which can be analyzed with whole-brain statistics. Regions that show significant pain-related activation would then be used as ROIs for subsequent tests of analgesic-related reductions. In this case, ROIs are based on the same subjects within the same scanner; one can even identify subject-specific ROIs if the localizer task is designed properly [70,90] (i.e. a substantial number of trials per condition during the localizer run; see [113,175,183] for details). Researchers can also use *a priori* ROIs based on the relevant literature. In this case, ROIs may be defined based on brain atlases that are in standardized stereotaxic space, using either neuroanatomical boundaries or extracting data from a sphere or box placed at specific coordinates. There are many atlases, and researchers must be judicious
in their selection of ROIs, based on their aims. Alternatively, ROIs can be based on meta-
analyses, e.g. with automated term-based meta-analyses such as through Neurosynth
[235]). Regardless of how individual ROIs are selected, the rationale for such decisions must
be reported and should be carefully evaluated.

If a researcher does not have strong *a priori* hypotheses about specific ROIs, or
wants to conduct whole brain analyses, they must use multiple comparisons corrections to
account for the number of tests and adjust p-value thresholds accordingly. Several
appropriate methods exist. For example, family-wise error correction (FWE) and Bonferroni
correction set adjusted p-values by dividing the threshold by the number of tests. This leads
to a very stringent threshold. The false discovery rate (FDR), in contrast, is a more lenient
method to correct for multiple comparisons, which controls for a false-positive proportion
(i.e., the fraction of detected voxels that are false positives; which is an unobservable metric)
[22]. In FDR correction, a rate (q) between 0 and 1, is specified, which represents the
maximum FDR which will be tolerated on average. Next, uncorrected p-values of
suprathreshold voxels are ranked based on significance (from smallest/most significant to
largest/least significant). Next, the voxel (or cluster) whose probability is greater than its
ranking divided by the total number of tests, as corrected by a desired FDR (q-value) is
identified. This voxel’s p-value is set as the threshold for a given FDR [91]. Each software
package might implement corrections slightly differently, e.g. some approaches account for
the spatial contiguity of activation (cluster-wise correction), whereas others set thresholds
based on voxel-wise values alone. Cluster correction is now advised [232] – especially for
FDR correction [44] – although there are notable differences in how this is implemented in
various packages [73] and this is an active area of debate in the broader neuroimaging
community [73,184]. Comparisons across these approaches have been discussed
elsewhere and we advise interested researchers to consult these papers for more thorough
treatment of considerations involved in methods for multiple comparison corrections
[24,73,151,152,172,232].

Another form of multiple comparisons correction is to perform non-parametric tests,
such as permutation testing, which do not make parametric assumptions about the data. Permutation testing resamples data and randomizes the assignment of observations. This process is repeated several times (usually >1000) to empirically determine a null distribution of the data, and to determine whether the observed differences are significant based on the data, and the alpha-value is set to minimize false-positives. A notable advantage of permutation testing is that it identifies the false-positive rate based on the data. However, permutation testing can be very computationally-intensive. For a comprehensive review, see [104].

Finally, we note that recent studies have shed light on an inflated rate of type-I errors in neuroimaging studies [73,105,110]. This indicates that there is an abnormally large number of significant findings that are reported in these studies – implying that negative results are not being reported (the so-called ‘file-drawer issue’, where negative results are not published, and that only results that meet statistical significance are reported, despite a large number of statistical tests performed (p-hacking). The field has become increasingly aware of and vigilant against such practices, as they can hinder true progress, and perpetuate false-positives.

4.5 Multivoxel pattern analyses.
The analyses we reviewed above are all referred to as mass univariate statistical tests, since each voxel is modeled as a single outcome in a statistical test, and multiple tests are conducted. In multivariate analyses, this relationship is switched – multiple voxels are modeled together rather than individually. One of the first multivariate analysis approaches was partial least squares [156], which identifies patterns, or distributed networks of brain activity related to a construct (such as a behavioural task). This approach was used by Seminowicz and colleagues to investigate the neural underpinnings of pain-cognition interactions [199]. More recently, this approach was used to compare networks dynamics to increasing cognitive loads in patients with fibromyalgia compared to controls [41].
More recently, classifier-based multi-voxel pattern analysis, or MVPA have become increasingly popular in pain research (see [103] for a review). In MVPA, algorithms from computer science and machine learning identify patterns across voxels that relate to a single construct. For example, MVPA has been used to test whether voxels within the anterior cingulate discriminate physical pain from other stimuli [49,141,220,231], or to identify features that can detect chronic pain [10,100,138,213]. In MVPA, machine learning algorithms use a subset of the data to train a pattern of weights across voxels that discriminate between states (e.g. acute pain vs other modality) [162,192]. The pattern is then tested against data that were not included in the training set to determine whether it can reliably discriminate between states (e.g. predicting that the condition was acute pain). This is repeated iteratively with different sets of held-out data through “cross-validation”, which then allows researchers to assess how well the pattern predicts outcomes. MVPA is a robust method to recover information encoded in ensembles of voxels. There are different algorithms and approaches that can be used for MVPA, but these are outside the scope of this primer. For a review, please see: [47].

4.6 Functional connectivity analyses
Another way to move beyond the consideration of individual activation clusters, or “blobs”, is by acknowledging the network nature of brain activity through functional connectivity analyses. Functional connectivity refers to a family of techniques that examine temporal correlations in BOLD activity across brain regions (highly related to the rs-fMRI techniques reviewed above). Critically, despite the term “connectivity”, functional connectivity does not rely on anatomical connections, but on simple correlations between the time courses of different regions. Functional connectivity approaches can be divided into several classes: those that assume static relationships between regions over time (“static connectivity”), those that assume that relationships vary over time (“dynamic connectivity”), and those that identify the influence of neural regions on one another (which can provide directional inferences in connections; “effective connectivity”). The simplest static connectivity measure
is the seed-based analysis, in which researchers extract the timeseries from a single ROI (the “seed”) and test for correlations between that region’s activation and activation throughout the rest of the brain. The region’s activation is used as the predictor in a GLM, and this results in a map of the strength of the correlations at each voxel which can be threshold according to the mass univariate approach described above. For example, one study found that patients with TMD have abnormally stronger connectivity between the medial prefrontal cortex and the posterior cingulate cortex/precuneus. This strength of this abnormal connectivity was related to pain rumination in these patients [129].

A more complex approach to static connectivity is to use principles from graph theory [19,32,87,98,207]. In this framework, brain structures are called nodes, and the connections between them are called edges. A graph theoretical approach captures the structure of a network (a set of nodes) and allows researchers to identify a variety of metrics for the network elements, including the density of connections of the brain regions (node degree), the clustering of neighboring nodes, the distance need to travel between two regions (path length, or efficiency), which regions are highly connected to many other regions (i.e. hubs), and which hubs are crucial (centrality). Together, these can be used to map out networks in pain, and how such networks may be altered in chronic pain conditions. Studies indicate that graph theoretical analysis of rs-fMRI can discriminate between placebo responders and non-responders in osteoarthritis [102,214], and has potential in the development of translational biomarkers for pain [122]. Additionally, these graph theoretical approaches have been used to investigate brain network abnormalities in chronic pain disorders [124,136,150,153]. While the details of this graph theoretical approaches are beyond the scope of this review, we refer readers to technical reviews of graph theoretical approaches to neuroimaging [19,32,75,87,207].

Whereas static connectivity measures assume that regions vary over time but their correlations remain stable, dynamic connectivity approaches allow the correlations across regions to vary as well. Dynamic connectivity has have been growing in popularity within the pain community [128]. While the field of dynamic connectivity is still rather new [114], one
type of dynamic connectivity approach has existed for quite some time: the psychophysiological interaction (PPI) [84]. This is essentially an interaction analysis, whereby a researcher can ask whether a condition modulates the connectivity between two brain regions. For example, a PPI was used to demonstrate that there is higher PAG-rACC coupling under placebo analgesia than under a control condition [26]. PPI can also be used to examine whether altered connectivity is associated with a behavioural state. This approach has been used to demonstrate that increased vIPFC connectivity with amygdala and nucleus accumbens during controllable pain was associated with reduced anxiety [196]. In PPI analyses, the “dynamic” aspect is known and driven by the experimental design (e.g. placebo blocks versus control blocks). In other approaches, dynamics are more fluid. Sliding window analyses examine connectivity across time by binning the task into specific chunks of time; this approach was used to demonstrate that individuals with more dynamic connectivity between mPFC and PAG were more apt to spontaneously disengage from pain [130]. Finally, new approaches can also use purely data-driven approaches to identify moments when networks reconfigure. For example, state-based dynamic connectivity analyses use latent models to estimate whether brain networks remain stable or shift connections over time [191]. This approach revealed dissociations among brain networks during remifentanil administration (e.g. connectivity within networks associated with emotion remained stable, whereas connectivity within networks associated with pain reorganized as drug infusion increased).

In contrast to static and dynamic connectivity, effective connectivity tests the plausibility of directional brain networks models – that is, not only is the activity between two brain regions correlated, but the effect of a neural region on another [81,83,86,87]. This allows for stronger inferences about the flow of information in the brain [157]. Effective connectivity usually requires that a set of brain regions of interest be defined a priori, although some path analysis approaches such as voxel-wise multilevel mediation analysis [7,221], allow for whole brain searches. One such method, dynamic causal modelling (DCM) uses Bayesian statistics to adjudicate between physiologically plausible models of
interactions between brain regions (or nodes) [81,149]. Once a model has been selected, the algorithm can determine how a condition can modulate the connectivity strength (or edges) between the nodes that comprise the network. This method has been used to determine the brain networks underlying isosalient nociceptive and innocuous stimuli of other modalities [143]. Mediation analysis tests whether the relationships between two variables is significantly reduced when a third, intervening variable is taken into account. Mediation can be used as a form of effective connectivity by testing whether the relationship between a brain region and behavior, or the correlation between two brain regions, can be partially explained by connections with another region. For example, this method has been used to show connectivity between the ventral striatum and ventromedial prefrontal cortex mediate the effects of self-regulation on pain [233]. Another study used mediation analysis to show that functional connectivity between the lateral prefrontal cortex and the PAG during rectal distention in healthy controls and patients with ulcerative colitis were mediated by the medial prefrontal cortex [155]. Notably, such a relationship was not observed in patients with irritable bowel syndrome, although the difference between the two groups was not formally tested and cannot, therefore, be considered a group difference. Another example found that the relationship between structural gray matter abnormalities in the supplementary motor area and pain-related helplessness in temporomandibular disorders was mediated by corticofugal motor white matter tracts [195].

One might assume that these group-level analysis decisions are atheoretical with respect to pain mechanisms. This is not, however, entirely the case. As an example, one of the most basic historical debates in the field is whether pain is the product of "labeled lines" running from the periphery to pain specific areas in the brain, or rather as a function of temporal and spatial patterns of activity throughout the neuroaxis. Labelled line theories date back centuries [163] and are largely based on neuroanatomic studies of spinal pathways and lesion studies. This locationist approach led to statistical analyses biased toward clusters of activation, such as traditional univariate GLM analyses. Local interactions or spatial patterns within these clusters are obscured by spatial blurring. As such, these
analyses have traditionally led to more modular interpretations of neural processing (i.e. that single regions have single functions) – a framework more consistent with a specificity account of neural pain processing. However, this account has largely been unable to identify specific brain regions that are necessary and/or sufficient for the experience of pain. As described above, MVPA has been used to increase specificity by searching for spatial patterns associated with particular psychological experiences. Such an analysis technique operates on the belief that activation more specific to pain will be identified by searching for unique patterns of spatial distribution, clearly more consistent with a pattern account. Similarly, studies employing functional connectivity or changes in functional connectivity across different time epochs are likely to result in interpretations of pain as a function of temporal and spatial patterning, rather than the product of activation of any single specific pain center.

4.7 Model-based analyses

Another approach to analyse fMRI data is the use of models to identify brain regions or networks related to perceptual experiences and more abstract concepts, such as the hedonic value of pain. One approach to modelling perceptual (and thus subjective) experiences, percept-related fMRI, was developed by Porro and colleagues [187]. This method requires a continuous rating of a measure of interest for the duration of the trial. This point-by-point (or TR-by-TR) perceptual rating curve is used as a regressor of interest in the statistical model. This method has been used to investigate prickle [58], paradoxical heat [59], and heat pain intensity coding [12] in healthy participants. It has also been used to investigate brain activity to experimental stimuli in chronic pain conditions [13,135].

To investigate more complex behaviours, such as decision making, valuations, and social interactions, neuroimagers have adopted engineering modelling approaches. Theoretical computational models of such processes are developed to estimate underlying computations and predictions about neural signals [61,62,202,234], and the dynamic interactions between theoretical signals. These models can generate predictive time-series
that can be used as regressors in the statistical model to identify correlated brain activity, and identify regions underlying the process. Given the complexity of pain, such approaches may be appropriate for pain research. Indeed, several families of computational models have been used to investigate the neural computations related to pain, including estimating the subjective value of pain and its relationship to cognitive factors and modulation [74,133,203,209,210,218,228,229,236].

5. Interpreting Data and Making Inferences

The ability to observe changes in oxygenated blood flow throughout the brain while a subject is in a particular behavioural or cognitive state is extremely powerful. Interpreting these activations, however, can be complicated, and the inferences that can be drawn are affected by a number of factors. Here we discuss some of these factors, and the limitations they pose on the interpretability of data.

5.1 Inference

Given the high cost of neuroimaging studies, and the difficulty in recruiting chronic pain populations, most pain imaging studies tend to be cross-sectional. These studies are important and useful as they can provide information on mechanism. For example, if a brain activation is significantly correlated with pain intensity, one might conclude that the activity is pain-related. However, observed activation might reflect many other correlated mental processes, so control conditions and interpretations must be carefully considered prior to making such conclusions. For example, correlations with pain intensity could reflect processes involved in magnitude estimation [12], salience detection [115,140,167], defensive reactions [164] or other processes that accompany pain but are not specific to pain. One must ensure that the control condition is sufficient for making claims about differentiating between these processes. Furthermore, the exact relationship between BOLD activation and underlying neuronal activity is still an open area of investigation [25,107,108,176,205]. Brain activation is inherently correlational, and authors often
speculate on the biological basis of a finding. While activations observed are related to the stimulus, one cannot infer causation from fMRI alone. In the absence of a task or intervention that can directly test a brain region (e.g., eliciting a virtual lesion using transcranial magnetic stimulation), we can only infer associations, but not causal links. For example, higher DLPFC activation might be associated with increased placebo analgesia (i.e. reduced pain; [8,48,200]), but if one wants to test whether DLPFC causes placebo effects on pain, one must show that placebo effects are abolished when DLPFC activity is disrupted. Indeed, researchers have done exactly this [200], indicating that DLPFC causally contributes to placebo analgesia. While this limits the inferences that might be drawn from fMRI alone, it is also clear that activation-based fMRI studies help identify brain candidates for causal interventions, and thus their potential utility remains enormous.

A related inferential difficulty arises when group differences are observed between pain patients and healthy controls in cross sectional studies. Without additional data, it is impossible to determine whether these differences pre-existed the condition (and therefore represent pre-disposing or possibly causal factors) or whether they are related to the cumulative effects of living with a chronic condition. Nevertheless, such inferences are frequently implied. For example, in structural imaging, it is common to refer to observed group differences as “plasticity” or “changes”, implying (without support) that the differences represent cumulative effects of disease. Some groups have attempted to form stronger inferences by performing correlations with disease characteristics, such as pain intensity, duration, questionnaire data, and other metrics [57] but even when such correlations are found, it remains difficult to determine whether they are due to cumulative effects of pain or other long-term effects of chronic disease (e.g. mood and behavioural changes).

Furthermore, such inferences are still limited by their correlational nature. Modeling techniques, such as dynamic causal modeling [85] and mediation analysis [146] allow inferences about the direction of information flow in neural circuits during a mental process or across participants, but they are still limited in causal inferences as they rely on the statistical variance within a cross-sectional dataset. The strongest way to address this issue
is to perform longitudinal studies in healthy subjects and/or chronic pain patients – scanning subjects before and after a therapeutic intervention (e.g. [97,118,131,201]) or before the progression from acute to chronic pain (e.g. [14,101]). Although such designs pose challenges in terms of resources and time, they allow for stronger and more mechanistically informative inferences about temporal precedence.

5.2 Logical fallacies in neuroimaging

Reverse inference is the widespread tendency to ascribe a function to an observed pattern of activation (e.g. “we observed activation in the visual cortex, therefore the participant is probably looking at something”). This issue has been discussed in depth in other reviews [116,182] so will only be covered briefly here. Reverse inference is tempting and, to some extent necessary, as the ultimate goal of scanning is elucidating mechanism. Caution is required, however, as reverse inference relies on a logical error (“affirming the consequent”). Despite the fact that the logic underlying such inferences is flawed, the probability that they are true can still be high if (as in the example of the visual cortex above) the region in question is linked fairly exclusively with a particular function. In pain imaging, however, we frequently observe activation of regions (e.g. insular and anterior cingulate cortices) that are involved in a wide variety of functions [235]. As a practical example of how this might affect inferences in pain studies, there is a widespread tendency to interpret observed activations in the periaqueductal gray as an indication that descending modulation is taking place. While such inferences can be strengthened by corroborating evidence (e.g. correlation with reduced pain ratings), it is commonly forgotten that the PAG is involved in many functions also relevant to pain, including escape and avoidance responses [16], prediction coding [193], and emotion [16,31,197]. Given this ambiguity, it is critical that experiments be designed to include appropriate control conditions and supplementary measures to adjudicate between these competing accounts.

A second common error that affects inferences is the so-called “imager’s fallacy”: that a difference in significance implies a significant difference [63]. Three instances of this
fallacy in imaging studies are group comparisons, implied lateralization and sub-regional specialization. Researchers often conduct separate analyses for each group in a between-groups study, e.g. to examine activation within patients and activation within controls. In such cases, it is tempting to display separate thresholded group maps and draw qualitative comparisons based on the extent of activation in each group. However, one cannot assume that a region was differentially activated simply because it meets a threshold for significance in one group and not in the other: apparent differences may simply reflect near-threshold differences (e.g. the p-value for a particular cluster was 0.049 for one group and 0.051 for the other). In order to conclude that there was a group difference, the groups must be formally compared using an appropriate statistical test (i.e. an interaction analysis). Relatedly, when comparing a correlation between two groups where one group shows a significant correlation with a behavioural measure, but the second group does not have a significant correlation, it is tempting to conclude that there are group differences between these correlations. However, the correlations coefficients must be normalized using Fisher’s z-transformation [78], and then formally compared. A similar error is often made with respect to laterality claims: If significant activation is observed on one side of the brain, but not in the symmetric contralateral brain region, it is common to conclude that the effect in question is lateralized. Without formally comparing the activation between the two sides, however, there is no basis for ruling out that processing is bilateral. Similarly, it is common to use significant activations to draw inferences about specialization of sub-regions of a given structure. As an example, based on anatomical tracing and other evidence, the insula has often been divided into a posterior region associated with sensory input and an anterior region involved in more abstract components of pain perception (e.g. interoception) [50-52,67,161]. Neuroimaging findings can only support this functional gradient if formal statistical testing is performed to demonstrate such selective functional association [117].

6. Conclusion
In this primer, we reviewed fundamental considerations in the design, analysis, description, and evaluation of a pain neuroimaging experiment. Investigators must make many careful decisions when designing experiments, and manuscripts must be written so that reviewers and readers in the community can evaluate the work appropriately, whether or not the reviewer has expertise in imaging. We hope this primer provides a foundation for pain clinicians and trainees without advanced training in neuroimaging, and encourage interested potential investigators to continue to read about the unique considerations involved in fMRI experiments by consulting recent dialogues that address these issues and their implications for the broader neuroimaging community.
REFERENCES


[221] Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-
subcortical pathways mediating successful emotion regulation. Neuron 59:1037-
1050, 2008.
[222] Wager TD, Keller MC, Lacey SC, Jonides J. Increased sensitivity in neuroimaging
[223] Wager TD, Rilling JK, Smith EE, Sokolik A, Casey KL, Davidson RJ, Kosslyn SM,
Rose RM, Cohen JD. Placebo-induced changes in FMRI in the anticipation and
reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at
integrates information about salience into perceptual decisions about pain. J
Modulation of pain processing in hyperalgesia by cognitive demand. Neuroimage
[228] Wiech K, Tracey I. Pain, decisions, and actions: a motivational perspective. Front
Tracey I. Combining fMRI with a pharmacokinetic model to determine which brain
areas activated by painful stimulation are specifically modulated by remifentanil.


FIGURE CAPTIONS

Figure 1: Number of pain imaging studies since the advent of fMRI. These numbers were determined with a search on PubMed using the search terms: “(Pain or nocicept*) AND brain AND (fMRI OR MRI) NOT CT NOT PET NOT EEG NOT review”. The search was restricted to 5-year time windows. The left panel shows the number of publications per 5-year windows. The panel on the right shows the cumulative number of studies over the years. The dashed blue line and the blue point have been extrapolated based on the current number of studies published from 2015 to January 2017.

Figure 2: Stimulus Convolution. Here, a hypothetical pattern of stimuli is convolved with the hemodynamic response function (HRF). In the upper panel, an event-related design is convolved to a double-gamma function to model the HRF. In the lower panel, the convolution of a block design study is shown. A prolonged stimulus would, in theory, elicit repeated HRFs. These are summated to produce a regressor to produce an idealized model of brain activity correlated to the stimulus. Note that the peak of the HRF is delayed with respect to the onset of the stimulus, and the offshoot is delayed with respect to the offset of the stimulus.

Figure 3: fMRI Pre-processing Pipeline. This flowchart represents one common fMRI pre-processing pipeline. The inclusion/exclusion of each step, as well as the order of steps, may vary based on a study’s unique goals and analysis plan. (*) denotes steps requiring quality check, optional steps are shaded in gray, steps specific to rs-fMRI (but optional in task-based studies) have a dashed outline. Adapted from [186].