Translational Neuroimaging for Infant Pain



Simon Marchant

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

Abstract

Many newborns are hospitalised, where they undergo clinically essential, noxious procedures. Exposure to these noxious stimuli is known to have harmful effects on infants in both the short-term and the long-term, and there are analgesic and other non-pharmacological interventions which might mitigate these. Despite this, pain is often undertreated in infants. This is partly because of the difficulty in identifying and measuring pain in this population. Our inability to accurately identify and measure pain in this population makes it difficult to evaluate the effect of interventions, and difficult to identify which hospitalised infants would benefit from them.

Neuroimaging measures of nociception have the advantage that they directly measure the brain. They can avoid many of the external confounding factors which are inherent to other methods, such as those based on physiology or behaviour. One particularly promising neuroimaging method is electroencephalography, which can be used to measure noxious evoked potentials (electroencephalographic signals caused by noxious stimuli). In this thesis I address two limitations of noxious evoked potential measurement; I do so by developing new methods for analysis.

The first topic is artefact and I address this in Chapter 4. Artefact in neuroimaging is measurement noise due to non-neural signals. I develop an automated, reproducible artefact detection algorithm. This algorithm uses machine learning techniques to distinguish between 'clean' epochs and those containing artefact. It performs as well as experienced raters in a fraction of the time.

The second topic is clinical utility and I address this in Chapter 5. A measurement for research settings does not necessarily translate well into clinical settings. For best clinical utility it is important for a measurement to be easily interpretable. I explore methods for combining individual neuroimaging and physiological metrics into a single, easily interpretable measure.

The overall intention of this thesis is to improve the EEG analysis methodology used both in neuroimaging research of neonatal pain, and in clinical practice.

Acknowledgements

No thesis has properly started without thanking one's family and supervisors. So thankyou to my parents, for supporting my vocational wanderings and for always being interested in my work. Thankyou also to my supervisors Rebeccah & Caroline. Your undying enthusiasm and implacable optimism are, if occasionally unnerving in their boundless quantities, a joy to work with.

Just as important are friends and colleagues. As with anything worth doing, producing a thesis takes help from so many people, in so many little ways: from the helpful conversation on a videocall to the encouraging smile on a morning. Any attempt to name all of you would be woefully inadequate, so I will just say thankyou. I think I have put a little bit of all of you into this thesis.

Table of Contents

Glossary **1**

List of Figures **3**

List of Tables **5**

1 Introduction **6**
1.1 Motivation . . . 6
1.2 Problem statement . . . 7
1.3 Objectives . . . 8
1.4 Thesis outline . . . 9

2 Background **11**
2.1 A critical review of the infant pain literature . . . 12
2.1.1 A definition of pain . . . 12
2.1.2 The neural basis for nociception . . . 14
2.1.3 The history of neonatal pain research . . . 15
2.1.4 Prevalence of neonatal pain . . . 16
2.1.5 Differences to adult nociception . . . 16
2.1.6 The impact of pain on neonates . . . 17
2.1.7 Analgesics for neonates . . . 18
2.2 Essential methodology and terminology . . . 19
2.2.1 Electroencephalography . . . 19
2.2.2 Event-related potentials . . . 22
2.2.3 A template of noxious activity . . . 23
2.2.4 Biomarkers . . . 24
2.2.5 Validation of biomarkers . . . 26
2.3 The state of the art in infant pain assessment . . . 28
2.3.1 Behavioural methods . . . 28
2.3.2 Physiological methods . . . 30
2.3.3 Brain haemodynamics methods . . . 31
2.3.4 Electroencephalographic methods . . . 34
2.3.5 The limitations of noxious-evoked potentials . . . 35
2.3.6 Other datasets . . . 38

3 Datasets **39**
3.1 General dataset information . . . 39
3.2 Dataset 1 . . . 40
3.3 Dataset 2 . . . 42
3.4 Dataset 3 . . . 43
3.5 Dataset 4 . . . 44
3.6 Discussion . . . 45

4 An artefact detection method for single-channel event-related potentials **47**
4.1 Introduction . . . 48
4.1.1 Background . . . 48
4.1.2 Objective . . . 50
4.2 Methods . . . 50
4.2.1 Datasets and study design . . . 50
4.2.2 Data processing . . . 51
4.2.3 Group identification of artefact (ground truth) . . . 51
4.2.4 Feature selection . . . 52
4.2.5 Model selection and hyperparameter optimisation . . . 55
4.2.6 Impact of ground truth . . . 56
4.2.7 Performance evaluation in the test set (Dataset 1 test split) . . . 58
4.2.8 Performance evaluation in the validation set (Dataset 2) . . . 58
4.2.9 Feature importance . . . 59
4.3 Results . . . 59
4.3.1 Artefact in EEG data can be consistently identified by individual raters . . . 59
4.3.2 EEG features are altered by artefact . . . 59
4.3.3 Artefact detection can be automated using a machine learning classifier . . . 60
4.3.4 Application of the artefact detection algorithm in the test dataset 62
4.3.5 Application of the artefact detection algorithm in a clinical trial 63
4.4 Discussion . . . 68

5 A clinically usable measure of infant nociception **73**
5.1 Introduction . . . 74
5.1.1 Objective . . . 74
5.1.2 Biomarkers of neonatal noxious response . . . 75
5.1.3 A diagnostic biomarker . . . 77
5.1.4 A monitoring biomarker . . . 78
5.1.5 Validation . . . 78
5.1.6 Chapter outline . . . 81

5.2 Methods . . . 83
5.2.1 Datasets and study design . . . 83
5.2.2 Response magnitude estimate . . . 85
5.2.3 Experimental stimuli . . . 87
5.2.4 Fit normalisation . . . 87
5.2.5 Baseline standardisation . . . 88
5.2.6 Heartrate modulation . . . 89
5.2.7 Evaluation metrics . . . 92
5.3 Results . . . 94
5.3.1 Response magnitude estimate . . . 94
5.3.2 Experimental stimuli . . . 94
5.3.3 Fit normalisation . . . 96
5.3.4 Baseline standardisation . . . 97
5.3.5 Heartrate modulation . . . 99
5.3.6 Evaluation using an independent validation dataset . . . 99
5.4 Discussion . . . 103

6 Discussion **107**
6.1 Aims of this thesis . . . 107
6.2 Contribution to the state of the art . . . 109
6.3 Limitations and future work . . . 110
6.3.1 Pain and nociception . . . 110
6.3.2 Acute and continuous pain . . . 111
6.3.3 Other measurement modalities . . . 113
6.3.4 Regulatory burden of proof . . . 114
6.4 Implications and impact . . . 115

Appendices

A A data pipeline for event-related potential studies **119**
A.1 Introduction . . . 119
A.1.1 Pre-processing of event-related potential data . . . 119
A.1.2 The need for an alternative approach . . . 120
A.1.3 Objective . . . 123
A.2 Data . . . 123
A.3 Pipeline description . . . 123
A.3.1 General architecture . . . 123
A.3.2 Lab-specific data transformation . . . 124
A.3.3 EEG pre-processing . . . 125
A.3.4 Response metrics estimation . . . 126
A.3.5 Artefact detection and reporting . . . 127

A.3.6 User interface 127
A.4 Filter responses 131
A.5 Pipeline output visualisation 135
A.6 Discussion 139

B Publications **141**

C Funding Acknowledgements **143**

D STROBE Checklists **144**

Bibliography **147**

Glossary

Abbreviations

EEG: Electroencephalography.

ERP: Event Related Potential.

ASR: Artefact Subspace Reconstruction.

AUROC: Area Under the Receiver Operator Characteristic.

FIR: Finite Impulse Response.

GA: Gestational Age.

HRV: Heartrate Variability.

IASP: International Association for the Study of Pain.

ICA: Independent Component Analysis.

(f)MRI: (Functional) Magnetic Resonance Imaging.

MEG: Magnetoencephalography.

NHS: National Health Service (in the United Kingdom).

NICU: Neonatal Intensive Care Unit.

NIRS: Near-infrared spectroscopy.

PNA: Postnatal Age.

PMA: Postmenstrual Age.

QMS: Quality Management System.

RCT: Randomised Controlled Trial.

ROC: Receiver Operating Characteristic.

SOP: Standard Operating Procedure.

Infant terminology

Infant: A child incapable of any form of independence from its mother. (Law and Martin 2020) NB. The term is usually used to refer to a child under one year of age.

Neonate: An infant at any time during the first 28 days of life. (Law and Martin 2020) NB. The word is particularly applied to infants just born or in the first week of life.

Term: Neonate whose birth occurs from the beginning of the first day (260th) of the 38th week through the end of the last day of the 42nd week (294th day) following the onset of the last menstrual period. (Committee on Fetus and Newborn 2004)

Preterm: Neonate whose birth occurs through the end of the last day of the 37th week (259th day) following the onset of the last menstrual period. (Committee on Fetus and Newborn 2004)

Postterm: Neonate whose birth occurs from the beginning of the first day (295th) of the 43rd week following the onset of the last menstrual period. (Committee on Fetus and Newborn 2004)

Gestational age: Duration of gestation based on best obstetric estimate. (Committee on Fetus and Newborn 2004) NB. This estimate is often the time between the first day of the last menstrual period and birth.

Postnatal age: Time elapsed since birth. (Committee on Fetus and Newborn 2004)

Postmenstrual age: Gestational age + postnatal age. (Committee on Fetus and Newborn 2004)

List of Figures

2.1 The neural pain template identified in Hartley et al. (2017). 25

4.1 Results of grid search hyperparameter optimisation. 57
4.2 Artefact rating by individuals. 60
4.3 Feature importance. 61
4.4 Feature magnitudes for each age group. 64
4.5 The magnitude of the absolute feature value in clean and artefactual epochs of EEG activity. 65
4.6 Model performance for the artefact detector in the training set. . . . 66
4.7 Posterior probability against rater decisions. 67
4.8 Results for test and validation sets. 67

5.1 Types of validity in the COSMIN framework. 80
5.2 Raincloud plot showing response magnitude distributions. 86
5.3 Examples of response magnitude estimates. 89
5.4 Illustrative example of heart rate modulation. 91
5.5 Proportional modulation optimisation. 92
5.6 No modulation: noxious and non-noxious stimuli. 95
5.7 No modulation: noxious stimuli and background. 95
5.8 Train of averaged stimuli compared to long-run average. 96
5.9 Fit normalisation: noxious and non-noxious stimuli. 97
5.10 Fit normalisation: noxious stimuli and background. 97
5.11 Baseline standardisation: noxious and non-noxious stimuli. 98
5.12 Baseline standardisation: noxious and non-noxious stimuli. 98
5.13 Proportional heartrate modulation: noxious and non-noxious stimuli. 99
5.14 No modulation on an independent validation set. 100

5.15 Fit normalisation on an independent validation set. 101
5.16 Baseline standardisation on an independent validation set. 101
5.17 Heartrate modulation on an independent validation set. 102

A.1 Schematic of the EEG data pipeline. 128
A.2 EEG visualisation for an infant. 129
A.3 A screenshot of the pipeline's graphical user interface. 130
A.4 Grand averages of EEG epochs filtered using different filter designs and high-pass cut-off frequencies. 134
A.5 Raincloud plots of three pipeline outputs, when the processing pipeline has been applied. 136
A.6 Raincloud plots of three pipeline outputs, when the processing pipeline has been applied to data from the Petal RCT. 137
A.7 Applied to data from the Petal randomized controlled trial. 138

List of Tables

3.1 Dataset 1 characteristics 40

3.2 Number of epochs used for artefact identification for each combination of age group and stimulus type. 42

3.3 Dataset 2 characteristics 43

3.4 Dataset 3 characteristics 44

3.5 Dataset 4 characteristics 45

4.1 Performance of alternative classifiers. 56

4.2 Optimised hyperparameters for the classifier. 56

5.1 Summary of Chapter 5 datasets. 85

5.2 Performance summary of modulations. 102

A.1 Published EEG pre-processing pipelines. 121

D.1 STROBE Checklist: Chapter 4 145

D.2 STROBE Checklist: Chapter 4 146

1 Introduction

Contents

1.1 Motivation **6**
1.2 Problem statement **7**
1.3 Objectives **8**
1.4 Thesis outline **9**

1.1 Motivation

Many newborns are hospitalised, where they undergo procedures which are clinically essential but potentially painful (Carbajal et al. 2008). Almost every medical procedure is potentially a painful experience for newborns: every blood test, every vaccination and every prod with a device could be painful. Exposure to these noxious stimuli has harmful effects on infants in both the short-term and the long-term (Ranger et al. 2015; van den Bosch et al. 2015; Vinall et al. 2013; Williams and Lascelles 2020). Despite the harm from these procedures, infant pain is undertreated in clinical inpatient settings (American Academy of Pediatrics and Canadian Paediatric Society 2006; Balice-Bourgois et al. 2020; Ranger et al. 2015). This is partly because of the difficulty in identifying and measuring pain in infants (American Academy of Pediatrics and Canadian Paediatric Society 2006; Slater et al. 2020). Our inability to accurately identify and measure pain in this

population makes it difficult to evaluate the effect of analgesics, and difficult to identify which hospitalised infants would benefit from these analgesics.

We cannot measure pain directly. The gold standard for identifying pain in the adult population is verbal report (Davis et al. 2017; Raja et al. 2020), which is not possible in infants. In the absence of verbal report, current practice relies on behavioural and physiological measures which are indirectly related to pain (Gibbins et al. 2014; Glenzel et al. 2023). The pain experience originates in the brain (Tracey and Mantyh 2007), and so brain activity may provide objective, quantitative and useful measures. Previous work in infants has identified an electrical, noxious-evoked potential at the scalp which can discriminate between noxious stimuli and non-noxious stimuli (Slater et al. 2010a).

Measures based on this noxious-evoked potential can be used in research settings, whether in clinical trials or in more exploratory research. Alternatively there is scope for development to use them in clinical settings, such as in cotside monitoring. There are many requirements in order for a measurement to be useful in a practical sense, and these requirements differ depending on the context in which it will be used. This thesis develops methods for using noxious-evoked potentials in research and in clinical settings.

1.2 Problem statement

Measuring noxious-evoked event-related potentials in infants provides an opportunity to directly measure changes in brain activity generated by nociceptive input in a non-verbal population who cannot describe their pain experience. However, as with any measurement technique, the measurement of noxious-evoked potentials in infants has limitations. My work addresses two limitations of this technique.

1. **Artefact.** Non-neural noise - artefact - reduces the reliability of the measurement by reducing the signal-to-noise ratio of the measurement. Manual identification and removal of artefact is time-consuming and subjective.

2. **Utility.** Methods that are developed to answer research questions cannot necessarily be used in a clinical setting. For example, in clinical trials the

effects of inter-subject variability can be reduced by averaging many individuals within a population; this approach is not possible when calculating a metric to quantify the response in a single individual.

The overall intention of this thesis is to improve the electroencephalographic (EEG) analysis methodology used for assessment of infant pain.

1.3 Objectives

The objectives of this thesis directly relate to the two limitations described above: improving the detection of artefact and the utility of noxious-evoked EEG activity. I will describe each here.

The objective of chapter 4 is to create an automated method of detecting artefact in single-channel 1500 ms epochs of infant EEG. In order to do so I train a random forest classifier on features identified from previous literature. In the chapter I test the classification accuracy of the model, and the generalisability to a new dataset.

Three separate datasets are used for testing classification accuracy.

- Training set (Dataset 1, grouped and stratified for cross-validation).
- Held-out test set (1/6 of Dataset 1, grouped and stratified).
- Validation set (Dataset 2).

Chapter 4 research question: For a) a held-out test set of EEG epochs, and b) a separate validation set of EEG epochs, is the balanced classification accuracy greater than chance?

The objective of chapter 5 is to assess the known-groups validity of a set of noxious response scales, each of which is a diagnostic and monitoring biomarker. To make a measure which can be clinically useful, I create a set of clinical scales and validate them using known-groups validity, a common method for validation of clinical scales. The noxious response scales are modifications of the response magnitude estimate (Hartley et al. 2017). The modifications incorporate information about the measured magnitude of the neurological response, confidence in that measurement, and the physiological response. They are as follows.

- Response magnitude estimate
- Experimental stimulus averaging
- Fit normalisation
- Baseline standardisation
- Heartrate modulation

Two separate pairs of known groups are used.

- Heel lance (an acute somatic nociceptive skin-breaking clinical procedure) against control heel lance (an acute somatic non-nociceptive non-skin-breaking experimental procedure).
- Experimental stimuli (an acute somatic nociceptive non-skin-breaking experimental procedure) against background (no procedure).

Chapter 5 research question: For a) an acute somatic clinical procedure, and b) an acute somatic experimental procedure, is the difference (noxious minus non-noxious) in group-level mean of modified response magnitude estimates significantly greater than the group-level mean of unmodified response magnitude estimates?

1.4 Thesis outline

This thesis intends to improve the infant pain EEG analysis methodology used both in neonatal neuroimaging research and ultimately in clinical practice. I do this by addressing two limitations of noxious-evoked potential measures: artefact and utility.

Chapter 2 reviews the literature relevant to the challenges of infant pain research. It then describes the state of the art in neuroimaging measures of infant pain.

Chapter 3 describes the datasets used in the thesis.

Chapter 4 addresses the limitation of artefact using an automated, reproducible artefact detection algorithm. This algorithm uses machine learning techniques to estimate the amount of artefact in any given epoch of EEG, which allows us to distinguish between 'clean' epochs and those containing artefact.

Chapter 5 addresses the limitation of utility by developing an algorithm which calculates a single metric from the noxious-evoked response and the heartrate. This incorporates information about the measured magnitude of the neurological response, confidence in that measurement, and the physiological response.

2 Background

Contents

2.1 A critical review of the infant pain literature 12
2.1.1 A definition of pain 12
2.1.2 The neural basis for nociception 14
2.1.3 The history of neonatal pain research 15
2.1.4 Prevalence of neonatal pain 16
2.1.5 Differences to adult nociception 16
2.1.6 The impact of pain on neonates 17
2.1.7 Analgesics for neonates 18
2.2 Essential methodology and terminology 19
2.2.1 Electroencephalography 19
2.2.2 Event-related potentials 22
2.2.3 A template of noxious activity 23
2.2.4 Biomarkers 24
2.2.5 Validation of biomarkers 26
2.3 The state of the art in infant pain assessment 28
2.3.1 Behavioural methods 28
2.3.2 Physiological methods 30
2.3.3 Brain haemodynamics methods 31
2.3.4 Electroencephalographic methods 34
2.3.5 The limitations of noxious-evoked potentials 35
2.3.6 Other datasets 38

2.1 A critical review of the infant pain literature

2.1.1 A definition of pain

The concept of pain has been discussed for quite a long time (Galen, 216 A.D., cited Harris 2018). Currently the International Association for the Study of Pain (IASP) defines pain as "An unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" (Raja et al. 2020). This definition includes two points which are important to us in the study of pain. First, pain is a sensory *and* emotional experience: an experience of pain must consist of both a sensory (sometimes called sensory-discriminative) component and an unpleasant emotional (sometimes called affective) component. Second, pain is associated with, *or resembling that associated with*, actual or potential tissue damage: while pain has an important role in avoiding tissue damage, it does not necessarily originate from tissue damage.

On the other hand, nociception is a different concept, which IASP defines as "the neural process of encoding noxious stimuli". Noxious stimuli are "actually or potentially tissue-damaging events transduced and encoded by nociceptors" (Raja et al. 2020). Although nociception tends to induce pain, the two can be independent from one another (Baliki and Apkarian 2015). For example, nociception may not induce pain for reasons such as attention (Bushnell, Čeko, and Low 2013) or neuropathies such as Charcot foot (Sanders 2004). Conversely, pain may be experienced without being induced by noxious stimuli, for reasons such as pain expectation (Benedetti et al. 2005; Wiech et al. 2014) or neuropathic pain (Colloca et al. 2017). Nociception and pain are separate concepts and one does not necessarily imply the other. Nociception can be measured - for example using neurophysiological recordings. Without verbal report, pain cannot be measured. In a scenario where we are measuring nociception because of its relation to pain, this non-linearity between nociception and pain adds complexity to how these nociceptive signals can be interpreted. The methods which are used and developed in this work are metrics of activity within the nociceptive system, which are not direct measures of pain.

Nociception is a necessary component of nociceptive pain and so a metric relating to nociception is one which relates to a component of nociceptive pain.

There have been various conceptual models of the pain system proposed. First, the tripartite model (Melzack and Casey 1968) models pain as comprising 3 dimensions: sensory-discriminative, affective-motivational, cognitive-evaluative. Of these, the sensory-discriminative dimension of the pain system encodes the spatial, temporal and magnitude properties of noxious inputs. The affective-motivational dimension is the aspect that encodes the motivation to take an action following a noxious input. The cognitive-evaluative dimension is the a component of the pain system, which is involved in the evaluation of the interpretation of noxious stimuli, which can be modulated by aspects such as attention, expectation and early experience (Melzack and Casey 1968). Another proposed conceptual model is the homeostatic model (Craig 2003). This conceptualisation holds that pain is an interoceptive sensation with an innate motivational value, and maps the physiological condition of the body. The motivation-decision model (Fields 2018) says that pain is a learning and motivational signal centred around predictions, and it can modulate itself. The predictive coding model (Büchel et al. 2014) models pain as a statistical sensory inference, implemented by hierarchical Bayesian architecture. Finally, the reinforcement learning model (Seymour 2019) proposes that pain expectation is updated based on the next available prediction. One way in which these models differ is in the purpose that they ascribe to the sensation of pain. In the 'sensory-first' models pain is useful as an estimated sense of a real noxious event. In the 'motivation-first' models pain is useful as a control signal to minimise future harm. These models are not necessarily exclusive and may coexist with one another (Seymour and Mancini 2020). The full subject of pain itself is both complex and difficult to disentangle in nonverbal infants. In this thesis I am concerned with hospitalised infants who experience noxious clinical procedures (see Section 2.1.4) and so I study nociceptive pain. Nociceptive pain is 'pain that arises from actual or threatened damage to non-neural tissue and is due to the activation of nociceptors' (International Association for the Study of Pain 1994).

2.1.2 The neural basis for nociception

Nociceptive information travels from the site of reception to the brain (the "ascending nociceptive pathway") by two types of nerve fibre: Aδ fibres and C fibres. Group Aδ fibres are large diameter bundles of myelinated neurons which carry noxious information quickly. Aδ fibres mainly carry nociceptive information about mechanical or thermal stimuli. Group C nerve fibres are narrow bundles of unmyelinated neurons which carry noxious information more slowly. Almost all C fibres carry nociceptive information of some kind (although around 10% instead carry non-nociceptive thermal information) and most C fibres are polymodal, in that they respond to nociceptive signals regardless of source (mechanical, thermal, chemical or otherwise) (Julius and Basbaum 2001).

Neuronal fibres which carry nociceptive information *to* the brain are called ascending fibres, and are divided into two tracts which form parallel neuronal paths within the ascending nociceptive pathway. These paths remain distinct throughout the journey from peripheral sensory nerves through the spinal cord to the brainstem and forebrain. One of these, in lamina V of the spinal cord, ascends through the anterolateral part of the spinothalamic tract and projects to the thalamus. This tract mainly carries sensory discriminative aspects of pain (location, intensity and stimulus modality) (Price and Dubner 1977). The other, in laminas I & II of the spinal cord, ascends through the lateral part of the spinothalamic tract and projects to the thalamus, parabrachial, amygdala, and periaquaductal grey. This tract leads to brain networks involves in generating the affective aspects of pain, and the tract mainly carries aspects of pain regarding pain control (Siegel and Sapru 2006).

The pain system is not unidirectional, though. As pain information ascends to the brain it is regulated by the descending pain modulatory system. This network can both inhibit and facilitate pain and nociception by modulating ascending pain signals when they pass the dorsal horn of the spinal cord, the medulla, and the midbrain (Tracey and Mantyh 2007). The descending pain modulatory system changes the experience of pain based on internal and external factors such as attention (Tracey et al. 2002), expectation (Benedetti et al. 2005), and mood (Giesecke et al. 2005).

The descending pain modulatory system appears to be present in both adults and infants (Goksan et al. 2018).

2.1.3 The history of neonatal pain research

Before the start of the 20th century the scientific consensus (at least in the Western world) appeared to be that newborns were incapable of feeling pain. It was only with the publication of the study "Pain and its effects in the human neonate and fetus" in 1987 that scientific opinion definitively accepted the idea that newborn infants were able to feel pain (Anand and Hickey 1987).

Early medical science had little interest in studying pain in newborns, partly because there was limited evidence to suggest that newborn infants had the capacity to experience pain (Anand and Hickey 1987). In fact, throughout the 19th century anaesthetic was often not used during surgical operations on neonates: when it was used, it was primarily to immobilise the patient. Rodkey and Riddell (2013) give four interdependent explanations for the assumption of infant insensitivity: the Darwinist belief that children and infants were lesser beings, a consensus that neonates were insufficiently matured for normal neurological reactions, the behaviourist belief that only psychological effects with observable behavioural effects were real, and an excessive reluctance to attribute to noxious response the observed responses to experimental noxious stimuli.

It was probably not, however, as simple as one scientific consensus immediately changing to its opposite. In the first half of the 20th century there was dissent in the neuroscience community over whether the interpretation of previous experiments was correct. By the second half of the 20th century, the matter had not been resolved and interest moved away from infant pain, with fewer studies conducted (Rodkey and Riddell 2013). Intra-operative anaesthesia was still not routinely used or recommended at this point. The 1987 study, and its vindication of some earlier thought, brought new interest to the field. Later studies confirmed the findings that human neonatal infants appear to react to, and feel, noxious stimuli.

It is now generally acknowledged that newborn infants can sense nociceptive inputs. We may not be able to know the pain experience of any individual infant, but there is a consensus that reducing pain in infants should be a priority (Roofthooft et al. 2014).

2.1.4 Prevalence of neonatal pain

Infants who are admitted to neonatal intensive care units (NICUs) are subject to many potentially-painful procedures every day. These procedures - performed as part of their routine clinical care - mostly consist of heel lances for blood samples and nasal and endotracheal suctioning, but can also include less common procedures such as cannulation, intubation and injections (Roofthooft et al. 2014).

Studies of the number of procedures estimate the average number of noxious exposures for infants in hospital as between 2 and 61 potentially-painful procedures a day (Carbajal et al. 2008; Cruz, Fernandes, and Oliveira 2016). In the UK, 1 in 7 newborns are admitted to hospital after birth, for an average of 7 days (Ougham and Modi 2017), which is predicted to increase with rising rates of premature birth (Blencowe et al. 2012).

2.1.5 Differences to adult nociception

Infants, being a non-verbal population, cannot describe their pain to us. This makes research into pain in infants difficult, as there is no "ground truth" on which to base our work. However, adults (usually) can describe the presence, magnitude and quality of their pain, which provides a ground truth which we can use to study it. This has led to a large amount of research into nociception and pain in adults. So the obvious question has to be asked: why not use adult methods on infants?

There is some suggestion that certain central nociceptive mechanisms may be similar between adults and infants. For example, the majority of the same brain regions are activated in both adult and infant nociception (Goksan et al. 2015) - with only the amygdala and orbitofrontal cortex activating in adults and not infants. The amygdala is thought to attach emotional significance to noxious inputs (Simons et al. 2014), and the orbitofrontal cortex is involved in anticipating future outcomes (Kahnt et al. 2010). This suggests that infants are too immature to contextualise a nociceptive

stimulus and evaluate a response (Goksan et al. 2015). The lack of activity in these regions could also be a false negative, and therefore the true overlap in adult and infant nociception remains to be discovered. Some studies have identified multivariate patterns of fMRI activity which are associated with subjective pain reports in adults (Wager et al. 2013; Woo et al. 2017). Some of these patterns have been shown to be present in both adults and infants when exposed to a mild noxious stimulus: those regions associated with ascending nociceptive afferents show similar activity in adults and infants, while infant and adult response activity diverged in regions such as the inferior parietal lobule and lateral occipital cortex (Duff et al. 2020). In addition to this, there are differences between adults in infants in the cortical activity evoked by noxious stimulus (Fabrizi et al. 2016). Taken together these studies show that, while there are similarities between adults and infants in pain-related brain activity, there are many aspects of nociception and possibly the pain experience that differ between the two populations. This makes it challenging to make conclusive assumptions about infant nociceptive response based on what we know of adult responses. The lack of knowledge in this area is important because exposure to pain in early life appears to impact infants' development, which I will discuss next.

2.1.6 The impact of pain on neonates

Noxious stimuli have acute effects on neonates such as changes in heart rate, respiratory rate, oxygen saturation, and motor activity. Other effects of interest to the clinical and research community are its long-term impacts on infant development.

Pain in early life appears to have several impacts on an infant's physiological and neurological development. Neonatal exposure to noxious stimuli is correlated with decreased white and grey matter maturation (Victoria and Murphy 2016) and poorer health outcomes later in life (Williams and Lascelles 2020).

The number of early-life exposures to noxious clinical procedures in the neonatal intensive care unit (NICU) is associated with poorer motor and cognitive development at 8 to 18 months' corrected age, independently of illness severity and medication administration (Giordano et al. 2023; Grunau et al. 2009). The number of exposures to noxious clinical procedures during early life is also associated with lower white

matter fractional anisotropy (a measure of white matter maturation) and with lower IQ test scores at 7 years (Vinall et al. 2014); also with white matter volume in the cerebellum at 7 and 10 years (Ranger et al. 2015; van den Bosch et al. 2015).

In addition, sensitivity to chronic pain is increased in children who were exposed to noxious stimuli as neonates. Neonatal exposure to noxious stimuli is related to increased sensitisation to noxious heat stimulus (Hermann et al. 2006), noxious cold stimulus (Vederhus et al. 2012), and noxious mechanical stimulus (Knaepen et al. 2013; Wollgarten-Hadamek et al. 2011) later in life, even though behavioural response to pain may be reduced (Sternberg et al. 2005).

As can be seen, exposure to noxious stimuli in early life has many long-term impacts on infants, including on their physiological and neurological development, on their motor and cognitive development, and on the development of their pain system. For this reason it might be beneficial to reduce infants' early-life pain exposure using analgesics.

2.1.7 Analgesics for neonates

Analgesic interventions - both pharmacological and non-pharmacological - are useful for relieving pain. Given the potential short and long term effects of exposure to pain as an infant, it would seem that the study and development of analgesic regimens for newborns would be important. However, pharmacological analgesics such as paracetamol and morphine have often been given to infants long before the first trials were performed, and even now the evidence base for their use has not caught up with their use (Baarslag et al. 2017; Foster, Taylor, and Spence 2017).

Analgesics have some risks, for example morphine can cause cardiorespiratory depression in infants (Hartley et al. 2018). Balancing risk with clinical benefit is important when administering interventions, and this is why it is important that analgesics are delivered in appropriate doses, backed by high-quality evidence.

The lack of evidence base for neonatal analgesic provision does not prevent these drugs from being used in neonates, and analgesics are widely prescribed in neonates despite not being licenced for use in this population (known as 'off-label use') (Costa et al. 2018; Geißler et al. 2020). However as this use is off-label, treatment

decisions must rely on clinical experience and extrapolation from other populations such as adults (Slater et al. 2020). This leads to a wide variety of different drug uses in clinical practice, and it is difficult to justify any particular approach to analgesic prescription (Flint et al. 2018). Having more evidence would allow for better analgesic dosage decisions to be made.

In order to gather more evidence on the best analgesic methods for neonates, more clinical trials on neonates are needed. It is particularly difficult to perform clinical trials in this population, and only around 2% of clinical studies worldwide are performed on neonates (Allegaert 2017). This is partly because good trial end points are objective (Higgins et al. 2019) and standardised (Davis et al. 2020) and it is difficult to achieve this using current methods (Slater et al. 2020). Developing and validating measures of noxious response in neonates would help to make it possible to run clinical trials on neonatal analgesics.

In conclusion, analgesics are often used clinically to relieve the experience of pain in neonates. However, there is little evidence available to underpin treatment decisions. By developing a high-quality measure of noxious response in newborn infants, we can make it possible to collect more evidence on which to base decisions about neonatal analgesia.

2.2 Essential methodology and terminology

2.2.1 Electroencephalography

An electroencephalogram (EEG) is electrical activity at the scalp, originating from brain structures, and measured by conductive electrodes. The existence of electrical voltages on the human scalp, and their connection to electrical activity in the brain, was first reported in 1924 by a scientist who placed electrodes on a subject's scalp and plotted the change in scalp voltages over time (Berger 1929; Wiedemann 1994). The field of electroencephalography was - rather unceremoniously - born.

Electroencephalography is possible because of local field potentials that are generated in the brain. These local field potentials are extracellular electrical field potentials which occur close to neurons (Amzica and Lopes da Silva 2017).

Neurons transfer electrical potential differences along the length of the neuron. At synapses, neurotransmitters and receptors modulate the permeability of specific ion channels in the postsynaptic membrane of the neuron. This generates an extracellular postsynaptic potential at the synapse (Amzica and Lopes da Silva 2017). In the cerebrum, the extracellular matrix is electrically conductive. Due to volume conduction, voltage changes within a neuron will cause very small voltage changes in the surrounding matrix. When a population of neurons are homogenously oriented and active in a coordinated way in time, the local field potentials that they generate sum together (Lopes da Silva 2013). The potentials induced by a neuronal population - and transmitted by volume conduction - can induce potential changes at the scalp. This makes it possible to use conductive electrodes to measure the sum of these potentials at the scalp: visible EEG can be seen from as little as 108 neurons (Olejniczak 2006).

Potential differences are always expressed as the difference between two points, and EEG measures the potential between several pairs of electrodes attached to the scalp. The recording of potential across two electrodes is called a 'channel'. This is often done using a 'reference electrode', which is not included in the analysis and to which every other electrode is compared. If we know the location and magnitude of all of the electrical charges within the brain, as well as the properties of the medium through which they propagate, then we can predict the potentials that we will measure at given points on the scalp (the 'forward problem'). However, if we know the distribution of potentials recorded at the scalp and wish to deduce the number, location and magnitude of sources then an analytical solution is not possible and assumptions must be made in order to make an estimate (the 'inverse problem', Wadman and Lopes da Silva 2017). Because of this inverse problem, EEG has poor spatial resolution compared to other neuroimaging modalities such as, for example, fMRI or functional near-infrared spectroscopy. However, EEG has good temporal resolution compared to other modalities. This is because the temporal dynamics of the signal being measured are fast, and the signal acquisition can also be fast. Temporal resolution compares favourably to functional magnetic resonance

imaging (fMRI), for example, as the temporal dynamics of synaptic activity are much faster than for haemodynamic activity (Luck 2014).

Any signal which is recorded over time can be deconstructed into a series of sine waves by Fourier Transform. EEG signals are often described in this way, using signal amplitudes within frequency ranges (called the 'frequency domain') instead of - or as well as - describing their signal amplitudes over time (called the 'time domain'). The frequencies present in an EEG recording are limited only by the methods used for recording, but certain bands of frequencies have been found to be particularly related to cortical activity. These bands have been given names (Tivadar and Murray 2019). From lower frequencies to higher frequencies, bands are delta (δ, 0.2-3.5 Hz), theta (θ, 4-7.5 Hz), alpha (α, 8-13 Hz), beta (β, 14-30 Hz), and gamma (γ, 30-70 Hz). Activity at different frequencies works in an integrated way, but activity at lower frequencies tends to engage large populations of neurons while activity at higher frequencies tends to involve relatively smaller, more discrete spatial regions (Lopes da Silva 2013).

The ease of use, low cost, and high temporal resolution of EEG compared to many other neuroimaging techniques has lead to its adoption in a number of clinical and research settings. EEG is widely used.

Some common uses for EEG in a clinical setting are epilepsy seizure detection, quantitative sleep monitoring, and anaesthetic status monitoring. For these applications, EEG usually uses several channels, in a standardised system known as the 10-20 system (Jasper 1958). Channels in the 10-20 system are referred to by two-part code which refers to the placement of the electrode. The first part is a letter (or two letters) corresponding to the dorsal/ventral location of the electrode. For example 'F' indicates 'frontal', ie. an electrode placed towards the ventral side of the head in the transverse plane; 'Fp' indicates 'frontal pole', ie. the front-most electrode on the head; 'C' indicates 'central', ie. an electrode placed centrally between the ventral and dorsal sides of the head. The second part is a number which indicates the side of the head ('1' for left and '2' for right) and distance from the centre, or 'z' which indicates an electrode placed centrally between the left and right sides of the head (Jasper 1958).

EEG is also used in neuroscience research. One application is in continuous monitoring of 'resting state' cerebral activity, for example in epilepsy research.

Another is in brain-computer interaction, where its direct measurement of potentials originating in the cortex is useful. It is also used in sensory research using responses evoked by experimental or clinical noxious stimuli. It is this last application which I focus on.

2.2.2 Event-related potentials

Event-related potentials (ERPs) are a type of EEG potential change which appears after an event of interest. The potential changes seen in ERPs are typically short and appear after the event in the time domain (Luck 2014). Because of this, certain techniques can be used to analyse this type of potential.

An EEG epoch is a short section of signal. By sectioning a longer signal into epochs, we can view each epoch of data independently. If we assume that the conditions at the time of each epoch are the same then we can treat these epochs in parallel, rather than in series. This is helpful, for example, if there is an event of interest within each epoch: rather than treating a recording as a single set of data with multiple interesting events inside it, we can treat the same recording as multiple sets of data, each containing an interesting event. Epoching data in this way allows us to compare events more easily (McLoughlin, Makeig, and Tsuang 2014).

In order to epoch the correct period of time within the signal, an EEG analyst must know at what time the event of interest occurred. The time at which the event occurred is typically labelled in the recording, and the synchronisation of the event timing with the recording timing is known as 'time-locking'. Doing so is not trivial, and the mechanism must be specific to the event being time-locked, but there are existing mechanisms for doing this (Lachaux et al. 1999; Worley et al. 2012).

Epoching data involves the removal of data in the time period surrounding the epoch. The point at which data is removed is chosen so that no data within the time period of interest is removed. However, an exact Fourier Transform requires an infinitely long signal in time. Any real-world calculations of a Fourier Transform necessarily involve approximation, where the precision of the approximation increases with signal duration (Stroud and Booth 2020). Frequency-domain calculations can be

distorted by epoching because of this. Epoching is therefore typically carried out after frequency-domain operations - such as filtering - have been performed (Luck 2014).

An evoked potential is an event-related potential which is caused by the event. This contrasts with an event-related potential which is any potential which is temporally related to an event of interest (whether caused by the event or not). Evoked potentials are a subject of some interest to the neuroscience community, for two main reasons. First, where an evoked potential can be identified it provides an indication of activity directly from the cerebrum in response to the event, which provides a more direct measurement than other forms of neuroimaging - such as fMRI - which rely on indirect measures such as blood perfusion. Second, multiple short epochs can be acquired by epoching EEG data which is time-locked to the event which stimulates the response. By averaging many epochs together, we can reduce the noise present in the recording (which tends to be Gaussian and distributed about zero) while leaving non-noise potentials intact. This allows us to reduce the noise present in an analysis (Luck 2014).

2.2.3 A template of noxious activity

Much of this thesis is concerned with a particular EEG measurement method, and so it is worthwhile to introduce this method in detail. This method has previously been called a 'template' of noxious neural activity.

An evoked response to noxious stimuli has previously been identified in infant EEG (Slater et al. 2010a), appearing in a group of 12 term-age infants after a clinically-required skin breaking heel lance but not after a non-noxious tactile control stimulus. Hartley et al. (2017) first characterised this particular 'template' representative of noxious response. To do this they used Principal Component Analysis (PCA) to identify a component which is higher in magnitude after a noxious than after a non-noxious stimuli in a group of 18 term-age infants. The component identified appears at the vertex of the head (the Cz channel on the standard 10-20 electrode system) between 400-700 ms after a noxious stimulus. The template increases in magnitude with increasing noxious stimulus intensity, and analgesics reduce the magnitude (Hartley et al. 2015, 2017). A diagram of the template is shown in Figure 2.1.

The method which I use to fit this template to new data consists of two steps. In the first step a modified Woody adaptive filter is applied. Woody filtering (Woody 1967) calculates cross-correlation between signal and template in order to identify the time shift with the maximum cross-correlation. Equation 2.1 shows cross-correlation Φ between signals f and g of length T and time shift ('jitter') $\Delta\tau$ (Woody 1967). In line with the method proposed in previous work (Cobo et al. 2021; Hartley et al. 2017) I modify the Woody adaptive filter by applying a single pass and limiting the possible jitter. I limit the possible jitter to a small range ($\pm$100ms) in order to ensure that the latency of the presumed event-related potential is biologically plausible (Cobo et al. 2021; Hartley et al. 2017). The Woody adaptive filter finds the jitter, $\Delta\tau$, corresponding to the maximum value of Φ.

$$\Phi = \frac{1}{T}\int_0^T f(t)g(t-\Delta\tau)dt \tag{2.1}$$

The second step is to fit the template to the recorded signal. In this thesis I do this by applying linear regression (equation 2.2) to the template (predictor variable) and new ERP data (response variable). The regression coefficient estimates represent the magnitude estimates of the event-related potential.

$$\bar{y} = \bar{\beta}\bar{x} + \bar{\epsilon} \tag{2.2}$$

2.2.4 Biomarkers

A biomarker is "a patient assessment that is not influenced by the patient's motivation or volition or a rater's judgement" (Walton et al. 2015). A biomarker is explicitly objective, and is *not* "a measure of how an individual feels, functions, or survives" (FDA-NIH Biomarker Working Group 2021). By this definition, most measures of pain in adults (based on verbal report) are not biomarkers: they are patient-reported outcome measures (PROs or PROMs) (FDA-NIH Biomarker Working Group 2021). Many measures of pain in infants currently used in clinical and research practice are also not biomarkers: they are observer-reported outcomes (ObsROs). As objective

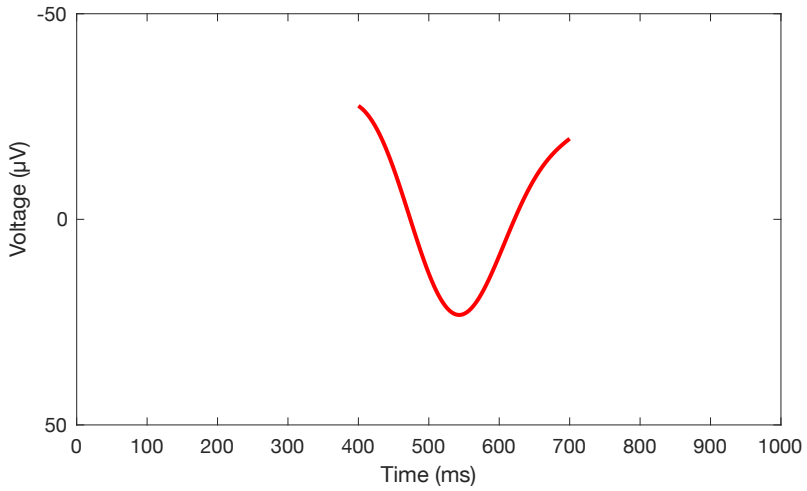


Figure 2.1: A visualization of the neural pain template identified in Hartley et al. (2017). The template is seen between approximately 400-700 ms after a noxious stimulus, although the exact delay varies with the characteristics of the stimulus and the individual. The magnitude of the signal shown here is arbitrary; for each epoch of interest the magnitude is adjusted to maximise fit between template and the observed signal.

measures of an infant's cortical activity, the EEG-derived methods used throughout this thesis are biomarkers (Baxter et al. 2023).

The FDA-NIH biomarker working group defines 7 categories of biomarker (FDA-NIH Biomarker Working Group 2016).

Susceptibility: A biomarker that indicates the potential for developing a disease or medical condition in an individual who does not currently have clinically apparent disease or the medical condition.

Diagnostic: A biomarker used to detect or confirm presence of a disease or condition of interest or to identify individuals with a subtype of the disease.

Monitoring: A biomarker measured repeatedly for assessing status of a disease or medical condition or for evidence of exposure to (or effect of) a medical product or an environmental agent.

Prognostic: A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest.

Predictive: A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent.

Response: A biomarker used to show that a biological response, potentially beneficial or harmful, has occurred in an individual who has been exposed to a medical product or an environmental agent. This response could be pharmacodynamic, or some other response used as a surrogate for a particular clinical benefit.

Safety: A biomarker measured before or after an exposure to a medical product or an environmental agent to indicate the likelihood, presence, or extent of toxicity as an adverse effect.

The methods used in this thesis can be diagnostic, monitoring, or response biomarkers depending on the context within which they are used. For example, in a clinical setting a measure of noxious response could be used to diagnose the presence of pain in an infant, or to monitor an infant's sensitivity to noxious stimuli over time. In a research setting, the same measure could be used to identify response to analgesics or anaesthetics, or to monitor an infant's sensitivity to noxious stimuli before, during, or after an intervention.

2.2.5 Validation of biomarkers

The COSMIN (consensus based standards for selection of health measurement instruments) Taxonomy of Measurement Properties identifies standardised definitions for the terminology of measurement properties (Mokkink et al. 2010). Other frameworks exist but COSMIN is used as it has comprehensive documentation (de Vet et al. 2011).

Reliability: the degree to which the measurement is free from measurement error.

Inter-rater reliability: the extent to which two or more raters, assessing independently, are consistent with each other.

Intra-rater reliability: the extent to which a single rater is consistent with themselves on two or more separate occasions. Also known as test-retest reliability.

Validity: the degree to which a measurement measures the construct it purports to measure.

Content validity: the degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured.

Construct validity: the degree to which the scores of a measurement instrument are consistent with hypotheses, e.g. with regard to internal relationships, relationships with scores of other instruments or differences between relevant groups.

Known-groups validity: A sub-type of construct validity. The degree to which an instrument discriminates between two groups known to differ on the construct of interest.

Criterion validity: the degree to which the scores of a measurement instrument are an adequate reflection of a gold standard.

Responsiveness: the ability of a measure to detect change over time in the construct to be measured. This can be a change in response to some external change, or caused by physiological changes within the patient.

Interpretability: the degree to which one can assign qualitative meaning (clinical or commonly understood connotations) to an instrument's quantitative score. This can be assessed by comparison to an external criterion representing an important measurement magnitude or

measurement change ('anchor approach'). Alternatively it can be assessed by using the distribution of a sample to identify an important change in the context of a population ('distribution approach').

Different properties can be more or less important for different types of biomarker. For example: due to the need to detect an individual's change over time, responsiveness is an important property for (as the name might suggest) response biomarkers. These properties do not map one-to-one onto common clinical measurement concepts such as accuracy, precision and reproducibility (Taktak et al. 2014), but are still valuable to bear in mind.

2.3 The state of the art in infant pain assessment

There are dozens of methods of measuring pain in neonates (Giordano et al. 2019; Glenzel et al. 2023), and they are proliferating. Evidently we have not yet found an ideal candidate, and so it is worthwhile investigating the broad categories and relative strengths and weaknesses of existing measures.

2.3.1 Behavioural methods

As pain is a subjective experience (Raja et al. 2020), it cannot be directly measured from behavioural response (Fitzgerald 2015). Behavioural response can be a surrogate measure and is often used for infant pain assessment in clinical settings (Gibbins et al. 2008, 2014; Hatfield and Ely 2015). Research into infants' behavioural response to pain has focussed mainly on three areas: crying, facial expression, and non-facial body movements (Hatfield and Ely 2015).

Attempts to use crying as an indicator of pain include studies on the latency and duration of crying following noxious stimulus (Franck 1986) and on the frequency content of sound produced by crying (Fuller 1991). Crying is a non-specific behavioural measure, however, and so there is so far little evidence that crying is a reliable indicator of pain in infants (Hatfield and Ely 2015).

Several patterns of facial activities are evoked in infants in response to noxious events, for example: brow bulge, eye squeeze, nasolabial furrow, open lips, vertical mouth stretch, horizontal mouth stretch, pursed lips, taut tongue, and chin quiver (Hatfield and Ely 2015). There is evidence from multiple studies that facial activity in infants differs after noxious stimulus (Hatfield and Ely 2015). There is contradictory evidence as to whether or not facial activities can be used to accurately quantify the intensity of nociceptive input, with some data suggesting that facial expression is correlated with noxious stimulus (Gibbins et al. 2008), and other data showing no correlation (Serpa et al. 2007). One possible reason for this is variability between raters in assessing the facial expression (Pillai Riddell and Craig 2007). In order to address this, several automated methods have been developed for classifying pain-related facial expressions from video, based on machine learning (Zamzmi et al. 2018). Another limitation of this method is that measures from facial expression are not specific to pain, for example facial expressions associated with painful experiences can also arise in response to tactile stimuli (Cornelissen et al. 2013; Green et al. 2019) and stressful experiences (Mörelius et al. 2006). Facial expressions can also be influenced by factors such as postnatal age (Schenk et al. 2019), gestational age at birth (Gibbins et al. 2008), and lethargy (Ranger, Johnston, and Anand 2007).

Body movements other than facial expression have also been studied in relation to noxious response. One study (Holsti et al. 2005) investigated the use of a multidimensional measure, the Newborn Individualized Developmental Care and Assessment Program (NIDCAP), in relation to heel lances (noxious stimuli) and comfort care (tactile stimuli). It found that 9 out of 26 behaviours differed between the two to a statistically significant extent, although it is not clear whether any measures differed to a clinically significant extent.

The need for manual rating of behaviour by care-giving staff takes time from other aspects of clinical care, limits the temporal resolution of these measures, and introduces the possibility of inter-rater error. Automated behaviour-identification methods have been developed (Brahnam et al. 2006; Heiderich, Leslie, and Guinsburg 2015; Zamzmi et al. 2016, 2018). However, these methods have several limitations

including an inability to function on partially-obscured, rapidly moving or poorly-lit infants (Heiderich et al. 2023). This has so far limited their practical utility.

There are some disadvantages to the use of behavioural measures in identifying pain. Perhaps most importantly, behaviour is modulated by other, non-noxious factors, and not all of these are easily measurable in themselves (Ranger, Johnston, and Anand 2007). For example nociceptive cortical activity becomes disassociated with an infants' behaviour when that infant is under stress (Jones et al. 2017). This creates confounds to the measurement which are difficult to account for. As Hatfield and Ely (2015) put it, "The inability of some neonates to mount a robust response due to gestational age or severity of illness and the variability in pain responses make identifying valid behavioral responses to pain difficult." (Hatfield and Ely 2015). Also, infants' noxious response varies with age (Green et al. 2019) and some very preterm infants show limited or no behavioural response to external stimuli such as noxious stimuli, due to immature and disorganised stress response (Johnston et al. 1995). As so many neonatal inpatients are premature infants, this population is an essential one to address in any measure of noxious response. Finally, it is not always possible to measure behaviour in infants who are hospital inpatients, due to practical limitations such as equipment preventing unobstructed view of the infant, or due to endotracheal intubation reducing an infant's ability to move freely. For these reasons, the sole use of behavioural measures for identifying and measuring noxious response in infants is unlikely to be successful.

While behavioural measures (including facial expressions) can be a surrogate measure of pain experience, they cannot directly measure pain, which is a subjective experience (Fitzgerald 2015; Raja et al. 2020; Slater et al. 2020). This does not prevent behavioural measures from being useful tools, but given the limitations described here it is worthwhile exploring other methods which can be used - either in conjunction with, or separately to, behavioural measures.

2.3.2 Physiological methods

Note that neurophysiological responses are not included here as they are discussed in Section 2.3.3.

The autonomic nervous system is linked to the pain system, and so it is reasonable to expect physiological responses to noxious stimuli. During and also immediately after noxious stimuli, heartrate tends to be raised (Gibbins et al. 2008; Holsti et al. 2005). Heart-rate variability (HRV) is a measure which is related to vagal tone and which is reduced in adults and infants experiencing noxious stimuli, fear or anxiety (Gee et al. 2017). High-frequency (0.15-0.4Hz) HRV is decreased in infants following noxious stimulus and this has been used to develop some heartrate-based measures (Butruille et al. 2015; Recher et al. 2021).

Sweating can be a physical reaction caused by the sympathetic nervous system in response to stressful stimuli such as pain, and this occurs in term-age and preterm neonates (Munsters et al. 2012). Physiological measures based on skin conductance have been shown to be related to behavioural pain measures after painful procedures (Gendras et al. 2021; Roué et al. 2018).

Physiological measures are correlated to behavioural measures to some extent (Ranger, Johnston, and Anand 2007), and in order to improve accuracy many established metrics for pain combine both physiological and behavioural aspects (Fitzgerald and Walker 2009; Roué et al. 2018; Salekin et al. 2021a). However, these multi-modal methods still have the same drawbacks as their constituent components, including limited stimulus specificity, confounding factors, and generalisability across sub-populations (Salekin et al. 2021a).

2.3.3 Brain haemodynamics methods

Neuroimaging measures of nociception have the advantage that they directly measure the source of pain: the brain. Behavioural and even physiological responses to external stimuli are all modulated by activity in the brain's cortex and subcortex (Siegel and Sapru 2006). This modulation can change with other, non-noxious factors such as stress, attention and previous experience. Metrics based on processes such as behaviour and physiology vary with these external factors in a way which is non-linear and cannot easily be controlled for; this introduces additional confounds in measures which cannot be easily estimated. Pain originates in the brain, and so measures of brain activity are as close as possible to the source of the phenomenon.

The first brain-derived measures of infant nociception used functional near-infrared spectroscopy (NIRS) to demonstrate a cortical response to noxious stimuli in infants (Bartocci et al. 2006; Slater 2006). NIRS has been shown to be able to identify cortical activation by comparison with simultaneous fMRI and PET imaging (Hock et al. 1997; Kleinschmidt et al. 1996). NIRS relies on the Beer-Lambert law (Beer 1852), by which light which passes through a solution is attenuated proportional to the absorptivity and concentration of the solute. Deep-red (near infrared) visual light has wavelengths of 700-1000nm. At these wavelengths oxygenated haemoglobin (HbO_2) and deoxygenated haemoglobin (HbR) have measurably different absorptivities. This allows an estimate of the concentration and oxygenation of blood in a region of the body by measuring the near-infrared light attenuation in that region. The technique is used to estimate oxygenation in regions of the human cranium in transcranial NIRS, and is particularly useful in neonates because near-infrared light can be transmitted through the entire head (Villringer et al. 1993).

Functional Magnetic Resonance Imaging (fMRI) is another brain-derived measure of nociception which has been used on infants (Duff et al. 2020). While 'standard' structural MRI measures the static distributions of different tissues within the brain, functional MRI measures dynamic changes in distribution of oxygenated blood (haemodynamics) in the brain. Similarly to NIRS, this is possible because of the different properties of oxygenated haemoglobin and deoxygenated haemoglobin: for fMRI it is their magnetic properties which differ. Haemodynamic changes occur in response to neuronal activity - called the blood oxygenation level dependent (BOLD) effect - and these can be measured by MRI because the local differences in concentrations of oxygenated and deoxygenated haemoglobin create measurable differences in local magnetic fields (Jenkinson and Chappell 2018). There are several protocols used for fMRI measures, most notably task-activation fMRI, resting-state fMRI, and stimulus-evoked fMRI.

- Task-activation fMRI measures haemodynamic activity when a specific task is performed. Activity during one task can be compared to activity in another. It is not possible to perform task-activation fMRI on infants as they cannot follow instructions.

- Resting-state fMRI measures haemodynamic activity in a brain at rest, and is used to correlate neuronal activity in different brain regions in order to identify communication between different anatomical brain regions (called functional connectivity).

- Stimulus-evoked fMRI measures haemodynamic activity when a sensory stimulus is presented. Like task-activation fMRI, activity during or after one stimulus can be compared to activity in another. This has also been used to identify noxious response in infants (Aspbury 2023; Baxter et al. 2021).

One interesting method for exploring brain activity related to subjective pain experience is in the use of multivariate fMRI models (Kragel et al. 2018). Previous work in fMRI on adults has developed multivariate models such as this to identify maps of brain activity associated with painful events (Marquand et al. 2010; Wager et al. 2013; Woo et al. 2017). These models, or 'signatures', appear to be highly correlated with the intensity of the nociceptive input and predictive of the intensity of pain verbally reported by an individual. One of these models (Wager et al. 2013) has also been shown to be present in infants (Duff et al. 2020) - see Section 2.1.5. These measures should be interpreted with care, as reverse inference based on brain region activation is difficult (Poldrack 2006). Even when we know that a cognitive process causes activation of a particular brain region or regions (forward inference), we cannot automatically infer that the same cognitive process is engaged based on the activation of those regions (reverse inference). The posterior likelihood that a cognitive process is engaged is influenced not just by the presence of related brain activity, but also by the prior likelihood of that cognitive process (Kragel et al. 2018; Poldrack 2006).

Other neuroimaging methods exist which are not reliant on brain haemodynamics. Magnetoencephalography (MEG) measures magnetic fields which are induced by electrical neuronal activity in the brain. While it has many advantages the required size and magnetic shielding makes it impractical in clinical settings (Hari and Puce 2023). Finally, EEG is the most commonly used technique for measuring noxious brain activity in infants.

2.3.4 Electroencephalographic methods

In contrast to the most common neuroimaging methods described here, EEG does not rely on haemodynamics. Instead it measures electrical potentials at the scalp. These potentials are induced by cortical and subcortical neuronal activity. This makes them the only common, clinically-feasible, brain-derived measure to be *directly* derived from cortical activity.

Noxious-evoked potentials are a particular type of evoked potential, where the pain system is activated by the evoking stimulus. There is evidence that these potentials are linked to subjective pain experience. In adults, noxious-evoked potentials are elicited by experimental mechanical noxious stimuli (Iannetti et al. 2013) which stimulate Aδ fibres (see Section 2.1.2). These noxious-evoked potentials consist of a negative-positive complex, and the magnitude of this potential is related to the self-reported perception of the pain intensity (Garcí-Larrea et al. 1997; Iannetti et al. 2005).

While behavioural and peripheral neurological responses to pain have been observed in newborn infants for some time, it was only relatively recently that a cortical response to noxious stimulus was identified in newborns by Slater et al. (2010). This was achieved by measuring evoked potentials to noxious stimuli in newborn infants. These noxious-evoked potentials are present in the 400-600 ms following a clinically-required heel lance (Slater et al. 2010a) but not a non-noxious control heel lance with the same tactile stimulus (only missing the skin-breaking lance). The magnitude of the potentials was increased in infants who had been born prematurely, compared to term-age infants of equivalent post-natal age (Slater et al. 2010c). This noxious activity template has since been validated in infants from 34-42 weeks' gestation in different research centres (Aspbury 2023; Cobo et al. 2022) and has been seen following heel lances (Slater et al. 2010a), immunisations (Cobo et al. 2021; Verriotis et al. 2015), and mildly noxious, non-skin-breaking, experimental stimulation (Hartley et al. 2015). This is the method described in section 2.2.3.

Aside from event-related potentials, other methods have been developed for assessing infant pain from EEG. Microstate analysis uses global field power, a description of overall brain activity. Local maxima of the global field power are

instants in time of the strongest overall field strength, and at these points microstate analysis considers topographies of the electric field to be discrete states of the EEG (Khanna et al. 2015). These individual topographies are analysed using, for example, spatial cluster analysis (Michel and Koenig 2018). Microstate analysis of resting-state EEG in adults has found differences between healthy controls and those with fibromyalgia (Michel and Koenig 2018) and chronic pain (González-Villar et al. 2020). In infants, microstates have been identified following a skin-breaking heel lance procedure which are consistent across infants and specific to the group having this procedure (Bucsea et al. 2023; Rupawala et al. 2023).

2.3.5 The limitations of noxious-evoked potentials

Like any measurement technique, the measurement of noxious evoked potentials has limitations. Here I will explain what I believe to be the two most important limitations. These are:-

1. Artefact
2. Utility

The first of these is artefact. An artefact is something observed in an investigation which is not naturally present, but occurs as a result of the investigative procedure. In the context of EEG this means that artefact is a noise source which is not of neural origin (Luck 2014). For example, this can be electromagnetic interference from nearby mains (often called 'line noise'), skin potentials, muscle activity, equipment-internal potentials which are amplified due to poor electrode fixation, or even signal loss due to faulty equipment. None of these effects are part of the signal of interest. Epochs of EEG which are contaminated by artefact have a higher magnitude of noise and so a reduced signal-to-noise ratio. This reduces the reliability of a measurement. In a research setting this potentially decreases the power of the research study which they are part of. In a clinical setting - where there is only a single patient and where there are fewer epochs to use - this could produce a higher or lower measured value than the reality. These effects are not desirable and so the standard approach is to identify and remove artefact wherever possible (Luck 2014).

EEG artefact identification and removal in adults is achieved by a number of methods. The most common is the set of techniques called Blind Source Separation, and the most common of those is Independent Component Analysis (ICA) (Luck 2014). This signal processing technique separates a signal into additive subcomponents, each of which is independent from the others. Using the technique in EEG assumes that different sources of EEG signal can be identified from their spatial distribution across channels. Once independent components have been identified, their magnitude within a given EEG epoch can be calculated. The source of each component can be identified manually by an expert rater or automatically by a machine learning method, and artefactual components can then be subtracted from each epoch. ICA is primarily a spatial method and so it requires a number of channels for effective source separation (Rejer and Gorski 2015).

Artefact is an especially common issue when measuring EEG in infants, for a number of reasons. First, by nature infants tend to move often and cannot understand - and will not comply with - instructions and so movement during recording is more common than in adults and children, who can understand and comply with instructions. Second, background EEG activity in infants is relatively high in amplitude compared to that of adults. Third, time that a research team can spend with an infant is very limited and so time taken to apply electrodes must be minimised; this requires limiting the number of electrodes applied, the time taken to reduce electrode impedance, or both. This leads to a more limited number of EEG channels and a higher channel impedance than would otherwise be the case. Acquiring data from a smaller number of EEG channels limits the amount of information which can be interpolated between channels, and acquiring data from a high-impedance system increases the impact of noise from external sources on the measured voltage. Lastly, it is particularly difficult to identify and remove eye-blink and eye-movement artefact in infants because the signal components are not as systematically reproducible as in adults (Fujioka et al. 2011).

Because of the difficulty in separating artefact from signal in infant EEG, the development of automated artefact identification algorithms is an area of interest in the paediatric neuroscience community. These methods allow researchers to apply

standardised, reproducible methods for artefact identification. However, the methods which have already been developed for use in infants require relatively large amounts of data, and it is not always possible to acquire this. For example, the number of channels and duration of recording is often limited by time constraints, and the number of stimuli which can be applied in ERP experiments can be limited in order to keep the infant comfortable. In particular, many noxious procedures would be unethical to perform unless clinically required, and so only a single epoch of EEG is available. An example is EEG that is recorded in response to a single clinically-necessary painful event, a heel lance performed for the purpose of blood sampling (Slater et al. 2010b). Studies on infants sometimes include EEG data that is recorded from a single recording electrode, on order to minimise disruption to the infants (Hartley et al. 2015; Slater et al. 2010b). There are no methods currently available which can be used on data with a single epoch and a single channel. This is the gap addressed by chapter 4.

Utility is the second limitation which my work addresses. In this context I use 'utility' to mean the ease of interpretation of a particular biomarker or metric. Ease of interpretation is an important characteristic for biomarkers to have when they are used in research settings, such when used as an outcome measure in a clinical trial (Prinsen et al. 2018). It is important that biomarkers that are used within a clinical setting are interpretable if they are to inform treatment decisions.

Currently, many measures of noxious response are conjugate measures, in that a single measurement reports multiple results. Users must interpret the meaning of these results in the context of their research question or diagnostic question. This limits the interpretability of this measure. Reporting a single scale which is a representation of the noxious response would improve utility by being simpler to interpret. In chapter 5 I create and test a number of scales of infant noxious response. Each scale is a modification of an ERP-based response magnitude estimate which has previously been identified (Hartley et al. 2017; Slater 2006). I evaluate each scale using known-groups validity.

2.3.6 Other datasets

The datasets used in this thesis consist of infant ERPs, many of which include noxious stimuli. All 3 datasets are part of a wider dataset collected primarily at the University of Oxford. Other datasets have been collected elsewhere. I will briefly summarise the availability of other datasets here.

The largest research group with relevant data is based in London. They have made their dataset available to researchers upon application (Jones et al. 2018). It consists of 20-channel EEG, heart rate, oxygen saturation and manually-identified behavioural scoring from 112 infants (29-47 weeks gestational age at study) immediately before and after three types of stimulus: clinical heel lance, control heel lance (vibrotactile stimulus), and auditory stimulus.

There are seven other geographic clusters of researchers which have published papers which indicate that they have collected infant EEG data in response to noxious stimuli. These potentially relevant datasets were collected in Ireland (Pavel et al. 2019); Norway, Sweden, and Finland (Castillo et al. 2024; Norman et al. 2008); Israel (Maimon et al. 2013); Egypt (Abdelsami Awad et al. 2018); USA (Maitre et al. 2017; Relland et al. 2024); and two distinct research groups in Canada in British Columbia (Askari et al. 2021) and Nova Scotia (Benoit et al. 2021). Each of these seven research groups have published at least one EEG study independently of the other research groups, suggesting there are at least seven discrete datasets globally. None of these datasets are publicly available.

There are two other publications on infant noxious ERPs which are not published by the research clusters above but do not contain usable datasets. One pre-print uses the publicly available UCL dataset (Talebi, Frounchi, and Tazehkand 2022). The other is a single-author paper from 1997 in which the EEG was recorded using an analogue machine with paper printouts (Scher 1997).

3 Datasets

Contents

3.1 General dataset information 39
3.2 Dataset 1 40
3.3 Dataset 2 42
3.4 Dataset 3 43
3.5 Dataset 4 44
3.6 Discussion 45

3.1 General dataset information

All of the datasets used in this thesis are secondary datasets, in that they were previously collected for prospective studies and are being re-used here. No new clinical data was collected for the purpose of this thesis.

All datasets used the same equipment for data collection. Electrophysiological activity was acquired from DC - 800 Hz with a sampling rate of 2 kHz using the SynAmps RT 64-channel headbox and amplifiers (Compumedics Neuroscan) and recorded using CURRYscan7 neuroimaging suite (Compumedics Neuroscan). EEG recording electrodes (Ambu Neuroline disposable Ag/AgCl cup electrodes) were positioned on the scalp according to the modified international 10-20 System, with reference and ground electrodes at Fz and the forehead respectively. EEG conductive paste (Elefix EEG paste, Nihon Kodhen) and EEG preparation gel (Nuprep gel,

D.O Weaver and Co.) was used to optimise contact, and impedances were < 10 kΩ. Studies used a varying number of channels, but always included Cz and this is the channel from which we take EEG for all chapters in this thesis.

In all datasets ethical approval was obtained from the National Research Ethics Service (references: 12/SC/0447, 11/LO/0350 and 19/LO/1085), and informed written parental consent was obtained before each study. Studies conformed to the standards of the Declaration of Helsinki and Good Clinical Practice guidelines.

3.2 Dataset 1

Dataset 1 is used in Chapter 4.

I selected participants from a database of data previously recorded at the John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. EEG from 160 infants was selected from that recorded for previous studies, published elsewhere (Cobo et al. 2021, 2022; Hartley et al. 2015, 2017; Schmidt Mellado et al. 2022; van der Vaart et al. 2019). Participant demographics are shown in table 3.1.

The inclusion criteria for an infant's data to be used within this dataset was as follows.

- 28-43 weeks postmenstrual age at time of recording.
- 23-42 weeks gestational age.
- Stimulus applied which was one of the following: auditory, visual, tactile, experimental noxious, control heel lance, heel lance, or manually-labelled periods of no external stimulation.

	Total	**Training set**	**Test set**
Sex (percent female)	46.5%	49.0%	33.3%
GA at birth (weeks)	35.0 (23.0 - 42.1)	35.9 (25.6 - 42.1)	34.6 (25.6 - 39.3)
PMA at test (weeks)	37.6 (28.0 - 42.7)	37.6 (28.0 - 42.7)	39.2 (38.7 - 39.7)
PNA at test (weeks)	2.6 (0.0 - 17.9)	1.71 (0.0 - 13.1)	4.62 (0.3 - 13.1)

Table 3.1: Dataset 1 demographics. Summary statistics are given as a percentage of participants, or as mean (range).

Different protocols were used for each type of stimulus.

Experimental stimuli were applied as follows. Resting state ("background") epochs were manually labelled during data collection as periods without external stimulation. Experimental auditory, visual, tactile and noxious stimuli were applied in trains of up to 30 stimuli, with an inter-stimulus interval of at least 10 seconds which was extended if the infant was unsettled, and automatically time-locked to the EEG. Auditory stimuli of 1 kHz frequency tones were applied using either speakers (X mini MAX II Portable Speakers, Xmi) placed 15 cm from the ears, 100 ms duration at a volume of 80 dB, or using headphones (Surpass Ltd, EMS Biomed), 4 ms duration and a volume of 70 dB nHL. Visual stimuli were single flashes of light (514 lumens, Lifelines Neuro) and tactile stimuli were applied with a modified tendon hammer which was used to gently tap the infant's heel, and were automatically time-locked to the EEG recording using a force transducer and event detection interface Worley et al. 2012. Experimental mild noxious stimuli (PinPrick, MRC Systems) of 64 mN or 128 mN were applied to the plantar surface of the heel. These were time-locked to the EEG using either an automated trigger (MRC Systems) or a high-speed video recording (220 frames per second (Hartley et al. 2015)).

Clinical stimuli were only applied once, as ethically they can only be applied as part of routine care. Heel lances were applied when necessary for blood sampling as part of the infant's clinical care. Vibrotactile stimuli were applied by performing a "control heel lance", where a heel lance was held against the infant's foot but rotated by 90 degrees so that when released the blade did not pierce the skin. Both types of event were automatically time-locked to EEG (Worley et al. 2012).

Where a stimulus was applied to a participant multiple times, only the first stimulus epoch was used in this analysis. This ensured that the effect of inter-subject variation was limited, as each infant contributed only a single EEG epoch for each stimulus type. This gave 495 unique EEG epochs.

The sample of epochs was then balanced for approximately equal numbers of epochs from each stimulus type. Balancing was achieved by removing epochs of types which were overrepresented in the data, and so a trade-off between the balance and size of the dataset was made subjectively. This balancing was completed while

Stimulus type	Premature 1	Premature 2	Term	Total
Background	18	9	31	58
Auditory	14	16	31	61
Visual	13	14	24	51
Tactile	13	13	27	53
Experimental noxious	10	11	39	60
Control lance	18	10	35	63
Heel lance	19	10	35	64
Total	105	83	222	410

Table 3.2: Number of included epochs for each combination of age group and stimulus type. Premature group 1. Infant born prematurely and epoch recorded less than 2 weeks from birth; Premature group 2. Infant born prematurely and epoch recorded more than 2 weeks from birth; Term: Infant born at term age and epoch recorded less than 2 weeks from birth.

blinded to the EEG traces, and the resulting distribution of data across ages and stimulus types is shown in table 3.2. This gave 410 unique EEG epochs.

3.3 Dataset 2

Dataset 2 is used in Chapters 4 and 5.

Dataset 2 consists of data from a previously conducted multi-site clinical trial, which involved EEG recordings during a clinically-required heel lance and a control heel lance. This randomised controlled trial (Hauck et al. 2024) investigated the effect of parental touch either before or after routine neonatal blood collection. Data was collected at two sites: the John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK; and the Royal Devon and Exeter Hospital, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK. Participant demographics are shown in table 3.3.

The inclusion criteria for this study were as follows.

- 35 weeks or more gestational age.
- Postnatal age 7 days or less.
- Require a heel lance as part of routine clinical care.

The exclusion criteria for this study were as follows.

Dataset 2 summary	
Sex (percent female)	39.5%
GA at birth (weeks)	38.3 (34.4 - 42.1)
PMA at test (weeks)	38.8 (35.4 - 42.9)
PNA at test (weeks)	0.4 (0.0 - 1.0)
Apgar at 1 minute (score)	8.4 (1 - 10)
Apgar at 5 minutes (score)	9.6 (5 - 10)
Apgar at 10 minutes (score)	9.9 (6 - 10)
Prior painful procedures (number)	4.6 (0 - 30)

Table 3.3: Dataset 2 characteristics, including demographics and factors which could influence pain perception. Summary statistics are given as a percentage of participants, or as mean (range). Acronyms are as follows (also see the Glossary, p1). GA: gestational age. PNA: post-natal age. Apgar: Apgar score, a measure of an infant's health immediately after birth (Apgar 2015).

- Intraventricular haemorrhage over grade II.
- Received analgesics or sedatives in 24 hours prior to study.
- Congenital malformation or genetic condition known to affect neurological development.
- Born to mothers who have a history of substance abuse.

The protocols for stimulus application were as follows. Heel lances were applied when necessary for blood sampling as part of the infant's clinical care. Vibrotactile stimuli were applied by performing a "control heel lance", where a heel lance was held against the infant's foot but rotated by 90 degrees so that when released the blade did not pierce the skin. Both types of event were automatically time-locked to EEG (Worley et al. 2012).

3.4 Dataset 3

Dataset 3 is used in Chapter 5.

Participants were selected from a database of data previously recorded at the John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. EEG from 87 infants was selected from that recorded for previous studies, published

Dataset 3 summary	
Sex (percent female)	43.7%
GA at birth (weeks)	39.9 (36.6 - 42.4)
PMA at test (weeks)	40.3 (36.7 - 43.0)
PNA at test (weeks)	0.4 (0.0 - 0.9)
Apgar at 1 minute (score)	8.2 (2 - 10)
Apgar at 5 minutes (score)	9.6 (7 - 10)
Apgar at 10 minutes (score)	9.9 (9 - 10)
Prior painful procedures (number)	4.3 (0 - 16)

Table 3.4: Dataset 3 characteristics, including demographics and factors which could influence pain perception. Summary statistics are given as a percentage of participants, or as mean (range). Acronyms are as follows (also see the Glossary, p1). GA: gestational age. PNA: post-natal age. Apgar: Apgar score, a measure of an infant's health immediately after birth (Apgar 2015).

elsewhere (Cobo et al. 2021, 2022; Hartley et al. 2015, 2017; Schmidt Mellado et al. 2022; van der Vaart et al. 2019). Participant demographics are shown in table 3.4.

The inclusion criteria for an infant's data to be used within this dataset was as follows.

- 36 weeks or more gestational age.
- Both heel lance (noxious) and control heel lance (tactile) stimuli were applied.

Stimuli were only applied once, as ethically skin-breaking procedures can only be applied as part of routine care. Heel lances were applied when necessary for blood sampling as part of the infant's clinical care. Vibrotactile stimuli were applied by performing a "control heel lance", where a heel lance was held against the infant's foot but rotated by 90 degrees so that when released the blade did not pierce the skin. Both types of event were automatically time-locked to EEG (Worley et al. 2012).

3.5 Dataset 4

Dataset 4 is used in Chapter 5.

Participants were selected from a database of data previously recorded at the John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. EEG from 87 infants was selected from that recorded for previous studies, published

Dataset 4 summary	
Sex (percent female)	42.5%
GA at birth (weeks)	38.8 (34.0 - 41.4)
PMA at test (weeks)	40.8 (16.4 - 90.7)
PNA at test (weeks)	5.2 (0.0 - 50.1)
Apgar at 1 minute (score)	9.3 (7 - 10)
Apgar at 5 minutes (score)	9.9 (9 - 10)
Apgar at 10 minutes (score)	10.0 (10 - 10)
Prior painful procedures (number)	5.4 (0 - 20)

Table 3.5: Dataset 4 characteristics, including demographics and factors which could influence pain perception. Summary statistics are given as a percentage of participants, or as mean (range). Acronyms are as follows (also see the Glossary, p1). GA: gestational age. PNA: post-natal age. Apgar: Apgar score, a measure of an infant's health immediately after birth (Apgar 2015).

elsewhere (Cobo et al. 2021, 2022; Hartley et al. 2015, 2017; Schmidt Mellado et al. 2022; van der Vaart et al. 2019). Participant demographics are shown in table 3.5.

The inclusion criteria for an infant's data to be used within this dataset was as follows.

- 36 weeks or more gestational age.
- Both heel lance (noxious) and control heel lance (tactile) stimuli were applied.

Resting state ("background") epochs were manually labelled during data collection as periods without external stimulation. Experimental mild noxious stimuli (PinPrick, MRC Systems) of 64 mN or 128 mN were applied to the plantar surface of the heel. These were time-locked to the EEG using either an automated trigger (MRC Systems) or a high-speed video recording (220 frames per second (Hartley et al. 2015)). These were applied in trains of up to 30 stimuli, with an inter-stimulus interval of at least 10 seconds which was extended if the infant was unsettled, and automatically time-locked to the EEG. Both types of event were automatically time-locked to EEG (Worley et al. 2012).

3.6 Discussion

The datasets used in this thesis have several characteristics that make them suitable for the work which I carried out. I discuss a non-exhaustive selection of these

characteristics here. In the previous chapter I discuss other datasets which have previously been collected containing similar data. The majority of these datasets have not been made available to researchers outside the team which collected the data. One dataset has been made available to researchers upon request (Jones et al. 2018).

The datasets used include a large number of individuals: 160, 114 and 87 infants for datasets 1, 2 and 3 respectively. This compares favourably to the one available alternative, which contains 112 individuals (Jones et al. 2018). Increased sample sizes improve our estimates of the wider population as a whole (Bland 2015).

The datasets used include different ranges of different infant ages. For example, dataset 1 includes infants between 24 and 42.7 weeks gestational age at time of test. The available alternative includes slightly older infants (29 to 47 weeks gestational age at test). EEG features change rapidly around the prenatal period (Bourel-Ponchel et al. 2021), and so a dataset which includes a wider range of premature infants has the benefit of including a range of prematurities. On the other hand, dataset 3 includes only term-age infants, as the ERP upon which Chapter 5 builds has different morphology in premature infants (Cobo et al. 2021; van der Vaart et al. 2022).

Dataset 1 includes 6 different stimulus types (plus stimulus-free background), compared to the 3 stimulus types (heel lance, control heel lance, and auditory) in the dataset published by Jones et al. (2018). Additional types of stimulus help to ensure that the model developed in Chapter 4 will be generalisable, rather than specific to a small set of stimuli. Dataset 3 includes heel lance, control heel lance, and experimental noxious stimuli in addition to background. This addition of experimental noxious stimuli allows the averaging of multiple response magnitude estimates, which is tested in Chapter 5.

In addition to the above points, data were collected for the datasets used using a standardised protocol. This includes, for dataset 2, pre-registration of the protocol used (Cobo et al. 2022). This standardisation of data collection method is helpful compared to, for example, combining datasets from disparate research centres. Different centres may have different methods of collecting data, which may not be apparent from published information and could impact an analysis of the data.

4

An artefact detection method for single-channel event-related potentials

Contents

4.1 Introduction 48
4.1.1 Background 48
4.1.2 Objective 50
4.2 Methods 50
4.2.1 Datasets and study design 50
4.2.2 Data processing 51
4.2.3 Group identification of artefact (ground truth) 51
4.2.4 Feature selection 52
4.2.5 Model selection and hyperparameter optimisation 55
4.2.6 Impact of ground truth 56
4.2.7 Performance evaluation in the test set (Dataset 1 test split) 58
4.2.8 Performance evaluation in the validation set (Dataset 2) . 58
4.2.9 Feature importance 59
4.3 Results 59
4.3.1 Artefact in EEG data can be consistently identified by individual raters 59
4.3.2 EEG features are altered by artefact 59
4.3.3 Artefact detection can be automated using a machine learning classifier 60
4.3.4 Application of the artefact detection algorithm in the test dataset 62
4.3.5 Application of the artefact detection algorithm in a clinical trial 63
4.4 Discussion 68

Related publication: Marchant et al. 2024
Part of this chapter was published in the above manuscript, under a CC-BY licence in the *Journal of Neural Engineering*. I am first author on the manuscript, and the content of this chapter is entirely my own. The work has been restructured and added to, to better fit a thesis format and to address examiner requests.

4.1 Introduction

4.1.1 Background

Electroencephalography (EEG) is a safe, non-invasive method of detecting and recording brain activity. It is commonly used both clinically and in a research context for assessing responses to sensory stimulation and for measuring background brain activity. A potential limitation of EEG is that the signals can frequently be contaminated by artefacts that can overwhelm the neurophysiological signal of interest (Luck 2014). These artefacts can arise from a variety of sources: examples include movement; detachment of the electrodes; electrical noise from inside or outside the body; and standing potentials and salt bridges on the scalp. Neonatal EEG, including stimulus-evoked EEG activity, is susceptible to movement artefact which can be difficult even for experienced users to accurately identify (Fujioka et al. 2011; Georgieva et al. 2020). This is of particular concern for studies of noxious stimuli, which evoke withdrawal reflexes (Hartley et al. 2015) and may suffer from contamination by stimulus-correlated motion.

Identification of artefact in EEG in both the clinical and research setting is usually performed by expert human review of the EEG signals. This is time consuming, requires expert input, and reduces the reproducibility of artefact detection, introducing a potential source of variability in results. To address this problem, a number of software-based EEG artefact detection algorithms have been developed (Chaumon, Bishop, and Busch 2015; Hatz et al. 2015; Mognon et al. 2011; Mullen et al. 2015; Nolan, Whelan, and Reilly 2010; Rosner 2012; Winkler, Haufe, and Tangermann 2011). The most common artefact detection techniques used in these

algorithms are Independent Component Analysis (ICA) and Artefact Subspace Reconstruction (ASR). ICA deconstructs the EEG signal into multiple component signals, based on the assumption that artefacts are generated independently from neural activity (Makeig et al. 1995). The obtained components (timeseries) are subsequently classified as being generated by a brain or non-brain source (Jung et al. 2000), using manual identification, thresholding (Delorme, Sejnowski, and Makeig 2007) or supervised machine learning (Pion-Tonachini, Kreutz-Delgado, and Makeig 2019), and non-brain components are removed from the original EEG timeseries (Jung et al. 2000). ASR relies on the identification of portions EEG that have high variance compared to a clean calibration set (Mullen et al. 2015). It can separate no more components than there are channels, and so cannot be applied to single-channel data (Kumaravel et al. 2021). These methods are much less effective at reducing signal-to-noise ratio with low numbers of channels (Rejer and Gorski 2015) and require long duration recordings, either from a single long recording or multiple short recordings from the same subject and session (Luck 2014). Alternative methods have also been developed, for example using the Extreme Standardised Deviate of the voltage amplitude to detect artefactual recordings (Smith et al. 2004). Recently, there has also been interest in automated methods for artefact recognition in infants, using ICA (Debnath et al. 2020; Fló et al. 2022), ASR (Kumaravel et al. 2022), and channel similarity (Gabard-Durnam et al. 2018). Most recently, Hermans et al. (2023) developed a machine learning model using semi-supervised methods to identify artefact in neonatal EEG. These methods have similar limitations to the ICA approach, and are optimal for long duration recordings from multiple electrodes.

The acquisition of EEG data in neonates presents some specific data analysis challenges related to the detection of artefacts. For example, many questions that are currently being addressed related to neonatal brain function and development involve short EEG recordings to identify event-related potentials (Hartley et al. 2015; Molfese 2000; Rupawala et al. 2023; Schmidt Mellado et al. 2022). In addition, low numbers of electrodes are often used due to small head size and infant compliance (Hartley et al. 2015; Molfese 2000; Rupawala et al. 2022; Schmidt Mellado et al. 2022). Further challenges also arise because the number of trials in infant studies

is often relatively small because infants' clinical vulnerability limits the number of stimuli which can be applied. In some cases, only a single trial of EEG activity may be available for each participant for ethical or practical reasons. For instance, studies looking at noxious-evoked brain activity often rely on recording the response to a clinically-required stimulus (e.g. a skin-breaking breaking blood test), which is usually only applied once during each recording period (Slater et al. 2010b).

4.1.2 Objective

An automated pre-defined artefact detection method that can identify artefacts in short-duration epochs of EEG activity recorded from a single electrode would improve research reproducibility, would reduce reliability on expert raters making EEG a more widely useable tool in neonates, and would be particularly useful in the context of recordings of event-related potentials (ERPs) in response to sensory stimulation. The aim of this study was to develop an automated artefact detector for ERP studies for neonatal EEG which fulfils these criteria and which was as reliable as a human rater.

Research question: For a) a held-out test set of EEG epochs, and b) a separate validation set of EEG epochs, is the balanced classification accuracy greater than chance?

4.2 Methods

4.2.1 Datasets and study design

Two internal pre-existing datasets were used to train, validate and test the model. I briefly introduce each dataset here; for further details on each dataset see Chapter 3. Dataset 1 consists of 410 unique EEG epochs from 160 infants 28-43 weeks postmenstrual age. These data were previously collected as part of ongoing studies in John Radcliffe Hospital (Oxford, UK) to investigate infants' responses to noxious and other sensory stimuli. Each epoch was 1.5 seconds long and contained one of 7 types of event at the 0.5 s mark: auditory, visual, tactile, vibrotactile, heel lance, mild noxious, or no event. I split this dataset into a training set (n=340 epochs), which was used to train and validate the model using cross-validation, and a test set

(n=70 epochs), which was used to test the final model on unseen data. This split was stratified by age and stimulus type and grouped by infant. As I envisage that the artefact detection model could be used in clinical trials using neonatal ERPs as an outcome measure, I also tested the model in a clinical trial dataset. Dataset 2 consists of 313 unique epochs from 108 infants 35-42 weeks postmenstrual age. These data were previously collected as part of Petal, a multi-site clinical trial conducted in the John Radcliffe Hospital (Oxford, UK) and the Royal Devon and Exeter Hospital (Exeter, UK) which investigates the impact of parental touch on infants' responses to a heel lance (Cobo et al. 2022). Each epoch was 1.5 seconds long and contained a heel lance, vibrotactile (sham heel lance) or no event at the 0.5 mark.

4.2.2 Data processing

For details of dataset collection and selection see Chapter 3. I carried out data analysis in MATLAB (Mathworks, version R2023a) and Python (Python Software Foundation, version 3.9.12). I band-pass filtered EEG using a non-causal windowed sinc FIR filter to 0.1 Hz - 70 Hz in line with recommended practice (Luck 2014). This is because a wide frequency range potentially allows a larger amount of information to be used for classification. A 49 Hz – 51 Hz band-stop filter excluded noise due to mains interference. This is because mains noise would be expected to be filtered in a real use-case (Luck 2014). From each participant I extracted a 1.5 second EEG epoch (-0.5 s to +1 s) around each stimulus of interest.

4.2.3 Group identification of artefact (ground truth)

To establish a ground truth for artefact, a group of 7 human raters independently rated each epoch as either containing artefact or not containing artefact. I designed an online questionnaire for this purpose. Epochs were presented in blocks of approximately 100, to reduce rater fatigue, and were randomised within each block. Raters were naïve to stimulus type and infant characteristics, in order to eliminate possible bias from the expectation of stimulus correlated motion. The ground truth (artefact versus no artefact) for a single epoch was determined by majority vote. All raters were paediatric neuroscience researchers from the University of Oxford, UK,

who had previous experience with identifying EEG artefact on single channel EEG epochs in a research setting. Three of the group were 'expert' raters (with more than 4 years' experience at EEG artefact detection) and 4 of the group were 'experienced' raters (with up to 4 years' experience). Expert raters had approximately 20, 12, and 6 years of experience and experienced raters had 3, 3, 2, and 1 years of experience. Only experience in a single-epoch, clinical trial or research setting is counted in this total as it is different to a clinical setting. This is due to different requirements for EEG to be rejected, and different information available at the multi-channel level.

The instructions supplied to raters in the questionnaire were as follows: "You are about to be shown 100 epochs of single-channel EEG amplitude data, recorded from term-age and premature infants within 4 weeks of birth. Please review each epoch and decide whether you would reject or include it in an analysis of evoked potential or evoked spectral analysis, based on the presence of artefact. We have included both stimulus-evoked data and data where there is no evoked activity. Where a stimulus has been applied, it is applied at the 0 second time point. You have been chosen because you routinely assess this type of data, and we would like you to treat this as clinical trial data. You will be looking at some short epochs (-0.5sec to +1.0sec). The y-axis of every plot is scaled to $\pm$ 150μV and minimal filtering has been applied (0.1-70Hz bandpass)."

In order to evaluate the rater performance against the classification performance, I used Cohen's Kappa to compare the performance of each pair of raters. An individual rater's average of their Cohen's Kappa with all other raters is an indication of how well their decisions are aligned with the rest of the group as a whole. This is preferable to using the balanced accuracy as a scoring mechanism, as it takes into account disagreements from the consensus within the group.

4.2.4 Feature selection

To develop a classification model which can identify artefacts in EEG epochs, I identified potential EEG features from existing literature which could predict artefact and which can be calculated for a single, short, single-channel epoch of data. These features are listed below. I only selected features that have been identified in the

literature to be associated with artefactual EEG activity. For each candidate feature, the publications from which I identified the feature are cited after the name.

- Amplitude change (Schmidt Mellado et al. 2022; Slater et al. 2010a). The maximum amplitude change over any 50ms period within an epoch.
- Mean local skewness (Winkler, Haufe, and Tangermann 2011; Winkler et al. 2014). The absolute local skewness in each 15ms segment. Here we take the mean of these local skews in an epoch.
- Temporal kurtosis (Mognon et al. 2011; Nolan, Whelan, and Reilly 2010). The kurtosis of each epoch.
- Amplitude variance (maximum local variance) (Mognon et al. 2011). Here we take the variance of an epoch's power.
- Derivatives (50/100/200/300 ms) - the sum of derivatives within a window (Chang et al. 2016). Here we use as features the maximum absolute sum of derivatives of the amplitude within a window, normalised to the amplitude variance of the epoch. Windows used are 50, 100, 200, 300 and 500ms, with each producing a separate feature in the model.
- Power < 0.5 Hz – the total power below 0.5 Hz (Luck 2014). Artefact often manifests at low frequencies, and the previous 0.1 Hz high-pass cut-off allows us to estimate the total low-frequency power in the epoch.
- Frequency band features (Winkler, Haufe, and Tangermann 2011; Winkler et al. 2014). These are applied to each EEG frequency band (delta 0.2-3.5 Hz, theta 4-7.5. Hz, alpha 8-13 Hz, beta 14-30 Hz, gamma 30-70 Hz (Tivadar and Murray 2019)) and also for the whole EEG spectrum - excluding <2 Hz, as EEG spectra tend to only follow a power law above 2 Hz (Buzsaki 2006). For each frequency band the following features are calculated.
  - Band power (Winkler, Haufe, and Tangermann 2011). This is an estimate of the whole-epoch power spectral density, calculated using a Fast Fourier Transform.

  - Lambda (Winkler, Haufe, and Tangermann 2011). This is the fitted gradient of plot of signal amplitude to log-frequency, from the equation $P = \lambda \log f + C$ where P is signal power, f is frequency, λ is lambda, and C is a constant (Donoghue et al. 2020). The amplitude is estimated with power spectral density and the fit is approximated using a least-squared curve fit.
  - Fit error – the maximum spectral deviation (Winkler, Haufe, and Tangermann 2011). This is the root-mean-square error between the line described by the equation above and the actual signal power spectral density.

- Fractal dimension (Accardo et al. 1997). The fractal dimension can be one measure of the complexity of a given signal (Ruiz de Miras 2016). It is defined by the equation $N = \epsilon^{-D}$ where N is the number of measurement units, ϵ is the scaling factor and D is the fractal dimension. Here we use two alternative estimates of the fractal dimension (Higuchi 1988; Katz 1988) and the features are the kurtosis of the fractal dimension using each method.

- Wavelets (Ghandeharion and Erfanian 2010; Orhan, Hekim, and Ozer 2011). Wavelet transforms decompose functions into series of waves, giving both spectral and temporal information. The Haar wavelet is one example, and the equation for Haar wavelet H at time t is shown below (Equation 4.1). Here we calculate the kurtosis of the Haar 1D wavelet transform (Haar 1910).

$$H(t) = \begin{cases} 1 & \text{where } 0 \leq t < 0.5 \\ -1 & \text{where } 0.5 \leq t < 1 \end{cases} \tag{4.1}$$

Feature selection was performed in the training dataset (n=340 epochs, approximately 5/6 of the available epochs in the Oxford Multimodal Dataset). For each candidate feature, the average absolute magnitude of the feature value for EEG traces classified as clean or artefactual was calculated. The features were plotted in order of p value from a binomial logit model comparing that feature's values for artefact and clean groups.

4.2.5 Model selection and hyperparameter optimisation

I performed model selection in the training dataset (n=340 epochs, approximately 5/6 of the available epochs in the Oxford Multimodal Dataset, stratified by age and stimulus type and grouped by infant). 71 epochs (21%) in the training set were rated artefact by the majority of raters. I assessed model accuracy in the training set using Leave-One-Subject-Out cross-validation.

Firstly, the most appropriate supervised-learning classifier was selected. I evaluated random forest, support vector machine (SVM), adaptive boosting (AdaBoost), and random under-sampling boosting (RUSBoost) models (Table 4.1). I compared performance of the different models using balanced accuracy (the average of sensitivity and specificity), and established statistical significance of differences between models using McNemar's Test for paired nominal data (McNemar 1947). All classifier types performed similarly and none had a statistically significant difference in accuracy. I therefore chose a random forest classifier for further model development - using scikitlearn v1.0.2 (Pedregosa et al. 2011). Random forest classifiers consist of an ensemble of decision tree classifiers. Each tree is built from a sample of the training dataset, drawn with replacement, and at each split in each tree the best split is found through a search of all feature values in a random subset of all features. The classification decision of the random forest is the popular vote of all trees in the ensemble (Breiman 2001; Cerulli 2023). These provide an easy and intuitive calculation of the posterior probability that an epoch contains artefact, and tend to perform well with multiple "weak learners", as is the case with multiple raters making subjective decisions such as this. Posterior probability - the probability that a given epoch would be rejected by the group of raters - is a useful metric for an investigator to review a classifier's decisions (this is discussed further in the Discussion).

Next I optimised hyperparameters. Observations were weighted by the certainty of group assessment (which could vary between all 7 raters agreeing and a 3-4 split). A grid search exhaustively searched all hyperparameter combinations within the set using Leave-One-Subject-Out cross-validation. The set of hyperparameters used is shown in Figure 4.1, with model performance at each hyperparameter value. The

Classifier type	Balanced accuracy	McNemar's p value
Random Forest	0.86	N/A
Support Vector Machine (SVM)	0.86	0.5
Adaboost with SVM	0.88	0.65
RUSboost with SVM	0.83	0.8

Table 4.1: Performance of alternative classifiers, when compared to a random forest model. Balanced accuracy for each classifier type is calculated on the training dataset. McNemar's comparison is used to give a likelihood that the classifier performs better than the random forest model. None of the alternative models perform statistically significantly better than a random forest (as indicated by p values above 0.05).

Hyperparameter	Optimised value
Max tree depth	10
Max features in tree	20%
Min samples per leaf	1
Min samples per split	5
Number of trees	100

Table 4.2: The set of hyperparameters used for training the random forest. These hyperparameter values were chosen based on maximum balanced accuracy achieved in a grid search optimisation.

grid search selected the set of classifier hyperparameters which gave the best (highest) f1 score weighted for class imbalance (i.e. each class's f1 score is weighted by its presence in the true data sample (Kelleher, MacNamee, and D'Arcy 2015). I used the selected hyperparameters (Table 4.2) to train an artefact identification model using the same weighting and grouping. I used Leave-One-Subject-Out cross-validation to evaluate model performance; I also assessed confusion matrix data, average Cohen's Kappa (by treating predictions in the same way as those of a rater) and balanced accuracy. The ground truth for each rater (including the artefact detector model) was calculated as the consensus decision for all other raters (i.e. each rater's decisions did not form part of the ground truth estimate for that rater). I calculated balanced accuracy for each rater from this ground truth.

4.2.6 Impact of ground truth

I trained the classifier using the popular vote of all 7 raters as the 'ground truth'. However, it is possible that this is not the most effective method for training the model: for example a more effective ground truth may be the popular vote of only

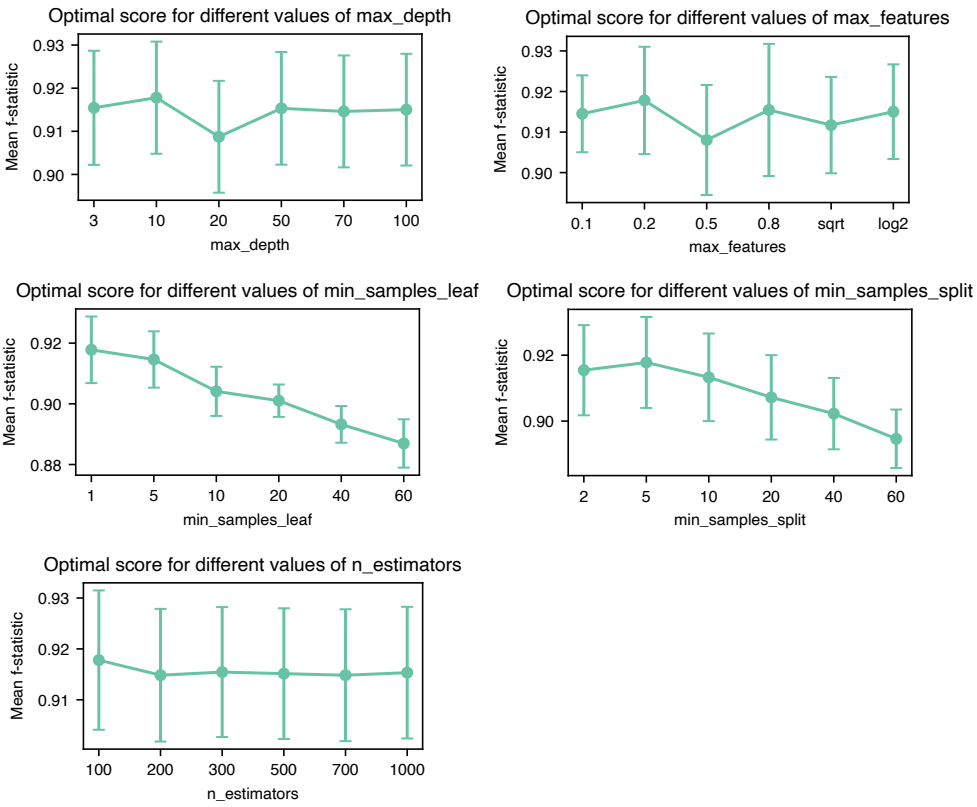


Figure 4.1: Results of grid search hyperparameter optimisation showing the impact of varying hyperparameters on the maximum achieved f-statistic. a) Mean f-statistic achieved by different values of the maximum tree depth. b) Mean f-statistic achieved by different values of the maximum features per tree. c) Mean f-statistic achieved by different values of the minimum samples per leaf. d) Mean f-statistic achieved by different values of the minimum samples per split. e) Mean f-statistic achieved by different values of the number of trees in the ensemble.

the 3 'expert' raters. These are the raters with more than 4 years of experience. From a data-driven perspective this is because these raters appear, from visual inspection, to agree with one another more often than other raters (Figure 4.2a). From a theory-driven perspective, this is because more experience may equate to more accurate decisions by these raters.

I tested the impact of changing the ground truth upon which the classifier was trained. I did this by training a random classifier on the training split from Dataset 1, with only the 3 'expert' raters contributing to the popular vote which formed the training target. For this I used the same optimised hyperparameters as previously. Comparison between the performance of the classifier using 7 raters and the classifier

using 3 raters was with McNemar's Test (McNemar 1947).

4.2.7 Performance evaluation in the test set (Dataset 1 test split)

I assessed the performance of the automated artefact detector on the independent held-out test data from the Oxford Multimodal dataset (n=70 epochs, approximately 1/6 of the Oxford Multimodal dataset). This dataset was kept separate from the training data and was not used for model selection. In the test set, 17 epochs (24%) were rated as artefact by the majority of raters. Confusion matrix data, average Cohen's Kappa (by treating predictions in the same way as those of a rater) and balanced accuracy were assessed to evaluate model performance.

I calculated ground truth for each rater (including the artefact detector model) as the consensus decision for all other raters (i.e. each rater's decisions did not form part of the ground truth estimate for that rater). I calculated balanced accuracy for each rater from this ground truth.

4.2.8 Performance evaluation in the validation set (Dataset 2)

A potential application of the artefact detection model is to use it in a clinical trial which uses ERPs as an outcome measure. Therefore, I additionally tested the artefact detection model in a previously conducted multi-site clinical trial, which involved EEG recordings during a clinically-required heel lance and a control heel lance. This randomised controlled trial (Cobo et al. 2022) investigated the effect of parental touch either before or after routine neonatal blood collection. Details on data acquisition and study design is reported in detail in Cobo et al. (2022). As part of the trial, epochs were manually labelled as artefact by two experienced raters. For the purpose of this work, I extracted epochs as for dataset 1, except filtering was performed at 0.5 - 30Hz to match the EEG to that assessed by manual raters. I compare classifier performance to the manual rejections made for the RCT.

In addition to accuracy statistics I calculated the impact of automated artefact identification on the signal-to-noise ratio (SNR) for the clinical trial. A measure of

SNR is calculated for each epoch as: $SNR = {}^{mag(ERP)}/_{std(baseline)}$ (Blum et al. 2019). I used only epochs with heel lance ERPs, for maximum expected ERP magnitude.

4.2.9 Feature importance

I assessed feature importance in Dataset 1's test dataset and in Dataset 2. For each feature in the trained model I calculated feature importance by feature permutation (using scikitlearn), randomly permuting each feature 100 times to determine the model's reliance on each feature. I chose this method over mean decrease in impurity as it can be computed on any model class, and so better generalises to other models.

4.3 Results

4.3.1 Artefact in EEG data can be consistently identified by individual raters

Seven human raters independently labelled a training dataset that contained 340 1.5-second epochs of EEG activity according to whether or not they contained artefacts. The EEG data included background activity (where no stimulus was applied) and responses to visual, auditory, tactile and noxious stimuli. The agreement between individual raters was 'fair to good', indicated by an average Cohen's Kappa of 0.51 (Figure 4.2a) (Fleiss 1981). The proportion of epochs rated as artefactual did not substantially differ across stimulus types (Figure 4.2b) or age groups (Figure 4.2c). As expected, agreement between raters related to their level of experience – Cohen's kappa between 'expert' raters (>4 years' experience) was 0.61 compared with 0.40 between 'experienced' raters (1-4 years' experience working with EEG, Figure 4.2a).

4.3.2 EEG features are altered by artefact

I identified from the literature a set of candidate EEG features that could be used to classify whether an epoch was considered to contain artefacts (see Methods). This included time-domain features (e.g. voltage amplitude variance, sum of derivatives within a window), frequency-domain features (e.g. frequency band power, maximum spectral deviation), and features using both the time and frequency domain (e.g. wavelet transform decompositions). For each candidate feature,

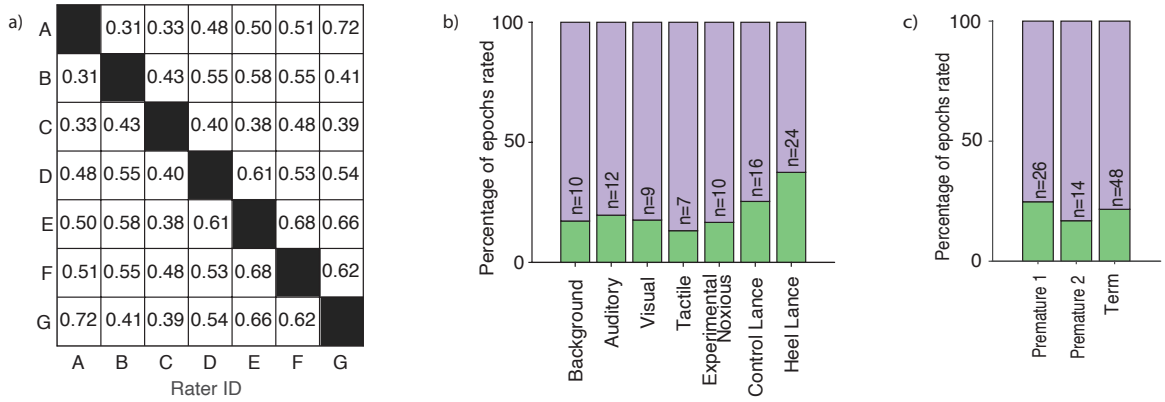


Figure 4.2: Artefact rating by individuals. a) Cohen's Kappa for each pair of raters. Each cell shows the Cohen's Kappa measure of agreement between a single pair of raters. Raters A-D are 'experienced' raters (with up to 4 years' experience) and E-G are 'expert' raters (with more than 4 years' experience). b) Percentage of epochs identified as artefact when split by stimulus modality. c) Percentage of epochs identified as artefact when split by age: Premature infants were born less than 37 weeks' gestation and term were born at 37 weeks' gestation or above. Premature group 1 are infants who had a postnatal age <2 weeks at the time of study, premature group 2 had a postnatal age > 2 weeks, all term-born infants had a postnatal age < 2 weeks at the time of study. In these bar charts, the lower (green) bar indicates epochs rated as artefact, while the higher (purple) bar indicates epochs rated as clean. Numbers on bars (i.e. n=X) indicate the absolute number of epochs represented by the green bar (i.e. the number of epochs containing artefact).

the absolute magnitude of the feature value for EEG traces classified as clean or artefactual is shown in Figure 4.5. This shows the features with the best potential for predicting artefact, although all features potentially contribute something to a combined classifier, even if not statistically significant by themselves. No features had statistically significant correlations with age (Figure 4.4). I used all features as part of the classification algorithm.

4.3.3 Artefact detection can be automated using a machine learning classifier

I have created a classification model to identify epochs of EEG activity with artefact compared to those without artefact (as defined by the consensus rater agreement), with four different models compared (see Methods). All models tested were comparable in terms of model performance ($p>0.05$, Table 4.1). The random forest model was used for subsequent analysis as the ensemble method used in a random forest allows for an easy and intuitive calculation of the posterior probability.

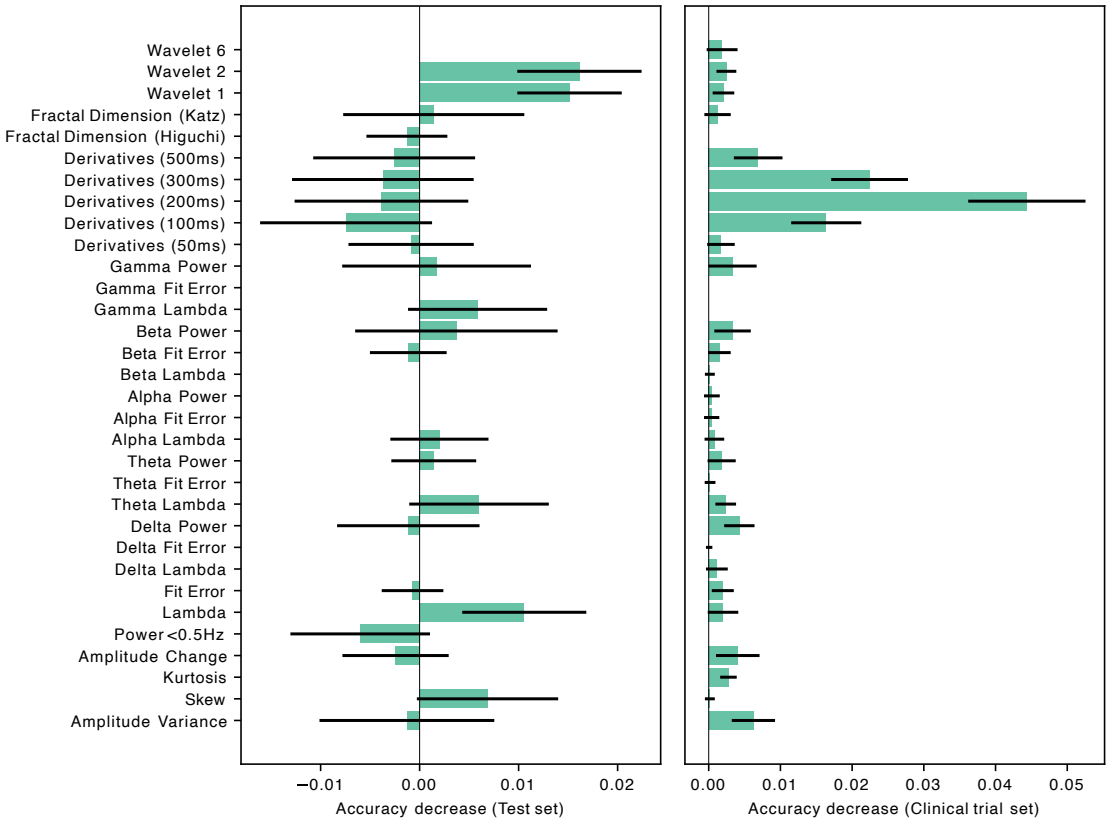


Figure 4.3: Feature importance when calculated on dataset 1 (left) and on dataset 2 (right). For each feature in the trained model, feature importance was calculated by feature permutation (using scikitlearn). For each feature this randomly permutes the relationship between feature value and epoch 100 times. The decrease in the model's accuracy when a feature is permuted is a measure of the importance of that feature. For each feature, the values shown here are for the mean and 95 percent Confidence Interval of the decrease in model accuracy for those 100 permutations. The importance of some features could not be identified, and in this case no data is provided in the graph.

This random forest model achieved a balanced accuracy of 0.86, a sensitivity of 0.80 and a specificity of 0.93 (assessed using cross-validation) in the training dataset (Figure 4.6a). Model performance was similar across different stimulus types (Figure 4.6b) and different age groups (Figure 4.6c).

The performance of the artefact detection classifier was better for those epochs with high rater agreement (Figure 4.6d). For example, when 6 of the 7 raters were consistent in the classification of the epoch the balanced accuracy was 0.95, the sensitivity 0.90, and the specificity 0.99 (n = 254 epochs). When there was no clear consensus among raters (i.e. when only 4 out of 7 raters reached a majority agreement) the model performance was lower (balanced accuracy = 0.68; sensitivity

= 0.55; specificity = 0.81; n = 64 epochs). As expected, higher posterior probabilities were generally correlated with higher rater agreement, and the curve approximates a logistic function (Figure 4.7).

When measuring agreement with manual raters, the artefact detection classifier had an average Cohen's Kappa of 0.57, which compares favourably with the consensus rater assessment. Figure 4.6e shows that the balanced accuracy of the classifier was comparable to that achieved by the individual raters, and overall had greater accuracy than 6 of the 7 individual raters. It performed better than the average balanced accuracy of both the experienced raters (0.80) and the expert raters (0.85).

I compared the performance of the classifier to a classifier trained only using the popular vote of the 3 'expert' raters (rather than the popular vote of all 7 raters). When compared, the classifier trained using only 3 raters did not perform significantly differently (balanced accuracy = 0.86; sensitivity = 0.75; specificity = 0.97; McNemar's p value = 0.65).

4.3.4 Application of the artefact detection algorithm in the test dataset

To estimate the prediction accuracy of the artefact detection classifier on unseen data, I applied the classifier to an independent held-out dataset. When tested on 70 held-out EEG epochs, the model had a balanced accuracy of 0.81 (sensitivity 0.65, specificity 0.98; Figure 4.8a). This compares well to an average balanced accuracy of 0.84 for all raters (calculated using the same leave-one-rater-out). Similar to the training data, comparison of posterior probability with rater agreement shows an approximation of a logistic function, suggesting that epochs which were easy for manual raters to classify were also easy for the detection algorithm to classify (Figure 4.7).

The classifier is optimised for maximum balanced accuracy, and the results given here are for that optimisation. It may be desirable to use a different balance of sensitivity and specificity: for example to prioritise a low false-negative rate. For this we can vary the probability threshold at which an epoch is classified as clean or artefact. The Receiver Operating Characteristic (ROC) curve shows how false

positive rate and true positive rate vary as the threshold is varied (Figure 4.8c). Feature importance for this dataset is shown in Figure 4.3 (left).

4.3.5 Application of the artefact detection algorithm in a clinical trial

The model is designed to provide an unbiased approach to rejecting data based on the presence of artefacts, and the previous test shows performance on held-out data from the same dataset. A method's ability to generalise to other data is particularly useful in the context of a clinical trial and so I also assessed performance on a second independent validation set that was collected as part of a clinical trial (Dataset 2). In the Petal trial, artefact was identified based on the consensus decision of two expert raters. Of the 316 epochs of data included in the trial (104 noxious heel lance epochs, 105 non-noxious control heel lance epochs and 107 background epochs with no stimulus), manual identification of artefacts in the original study led to 46 epochs being identified as artefactual. In comparison, when the automated artefact detection classifier was used, 103 data epochs were identified as artefactual. A confusion matrix shows the performance of the artefact detector on the novel clinical trial data: 0.78 balanced accuracy, 0.67 sensitivity, 0.88 specificity (Figure 4.8b).

The Receiver Operating Characteristic curve (Figure 4.8d) has an area under the curve of 0.89. This is similar to, and slightly lower than, that of the test set, which has an area under the curve of 0.96.

I calculated mean SNR for the group of heel lance epochs identified as clean by the detector (n=68, SNR=0.052), and for the group of heel lance epochs included in the original clinical trial (n=82, SNR=0.046). SNR using the automated method was 12% higher than using the expert team of the original trial.

Feature importance for this dataset is shown in Figure 4.3 (right), alongside that of the held-out dataset.

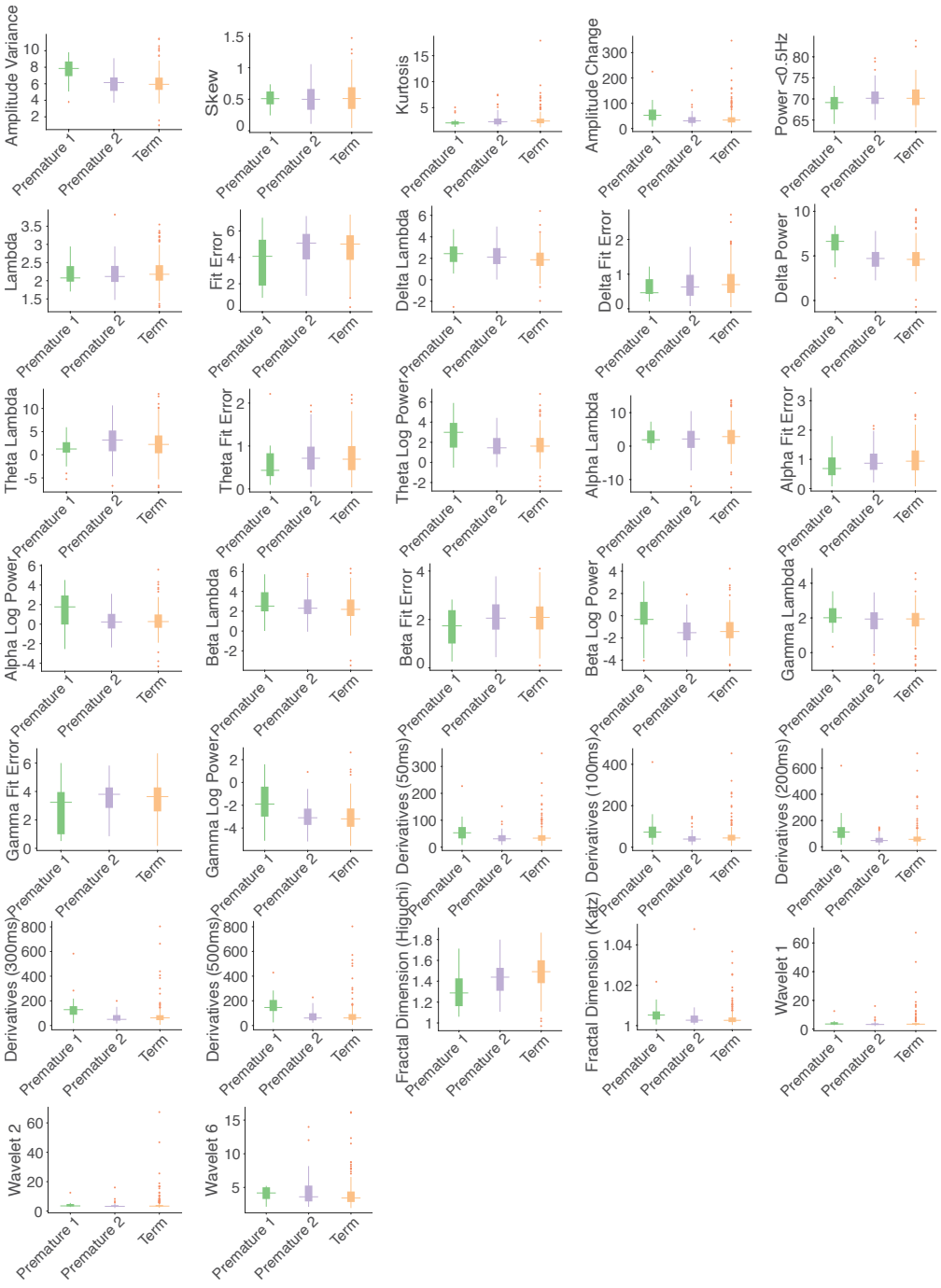


Figure 4.4: Feature magnitudes for each feature. For each feature, data is split by infant age group, where the group "premature 1" is infants born prematurely with postnatal age less than 2 weeks at time of test, and "premature 2" is infants born prematurely with postnatal age 2 weeks or above at time of test. No age groups significantly differed in feature value.

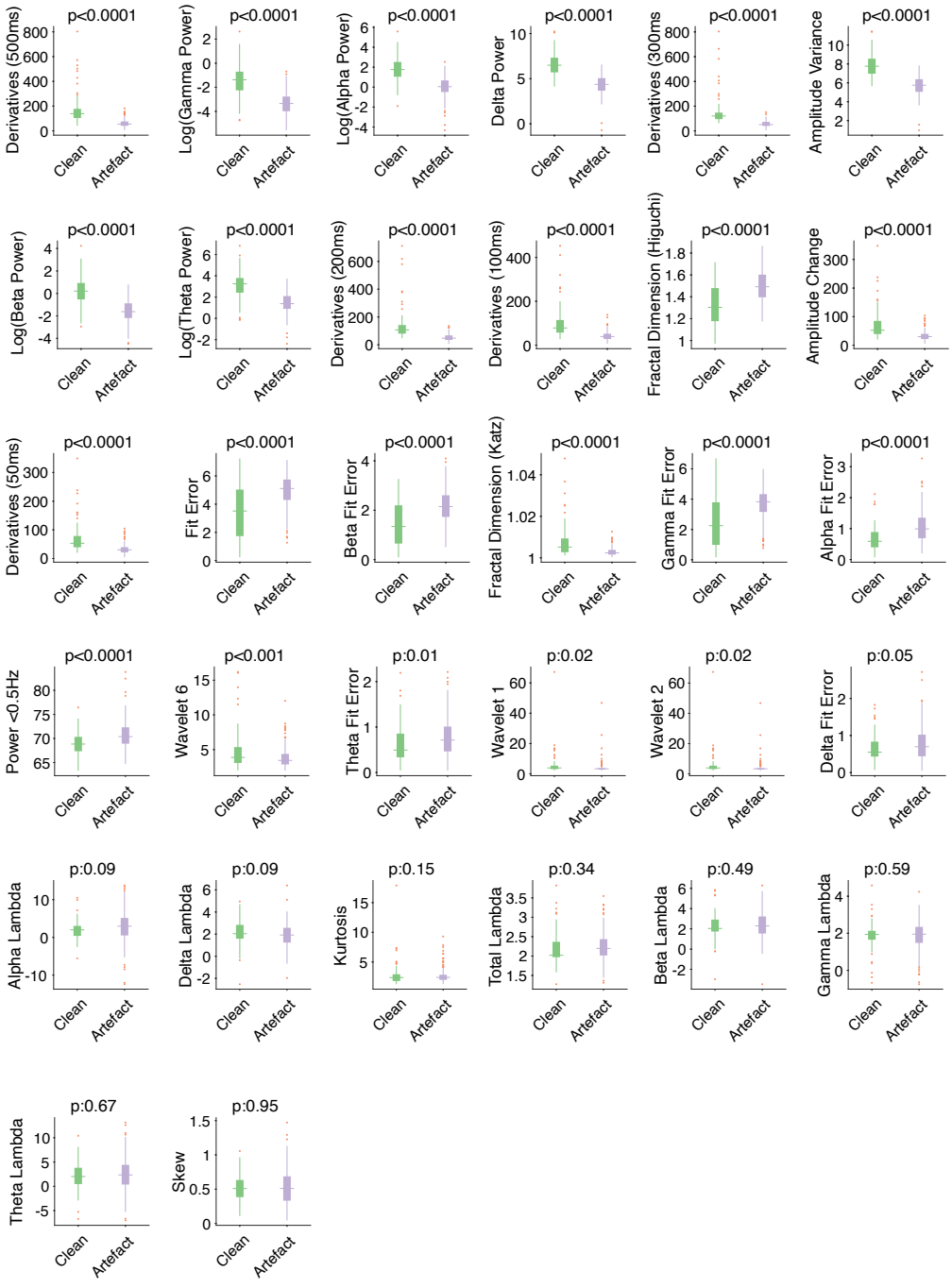


Figure 4.5: The magnitude of the absolute feature value in clean and artefactual epochs of EEG activity (as decided by majority decision from raters). Above each plot is the p value related to the null hypothesis that the data are drawn from the same population (as calculated by binomial logit model for that feature). Feature values are transformed into log values where appropriate for easier viewing - as shown in y-axis labels.

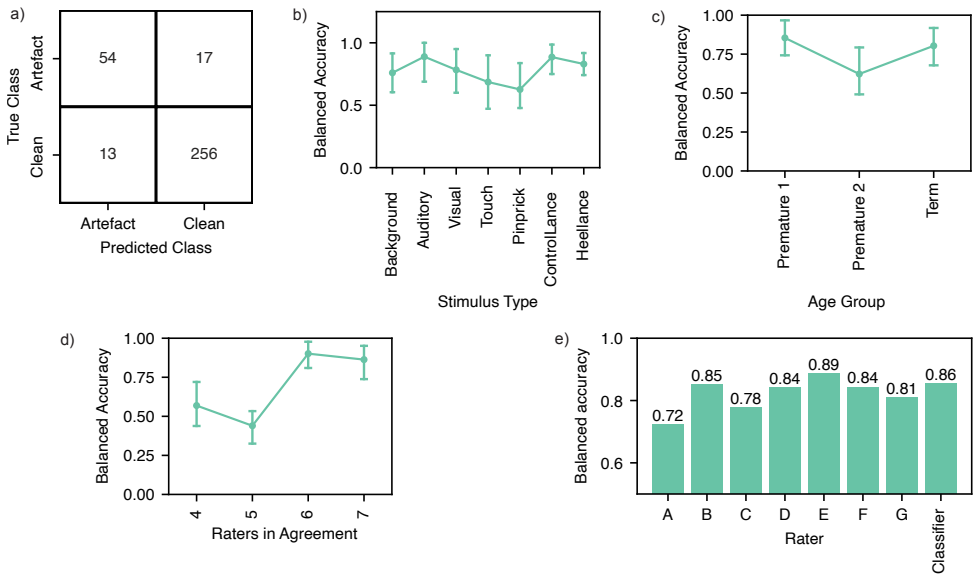


Figure 4.6: Model performance for the automated artefact detector in the training set. a) Confusion matrix for data in the training set (n=340). Numbers indicate the number of epochs classified correctly and incorrectly according to whether or not the epoch contained artefact. True class indicates the majority decision of the raters, predicted class indicates the class predicted by the automated artefact detector. From this the sensitivity (0.80), specificity (0.93) and balanced accuracy (0.86) can be derived. b) Classifier performance according to type of stimulation in each EEG epoch; c) Classifier performance according to age group, where the group "premature 1" is infants born prematurely with postnatal age less than 2 weeks at time of test, and "premature 2" is infants born prematurely with postnatal age 2 weeks or above at time of test; d) Classifier performance on subsets of the data split according to the number of raters in agreement with the consensus; e) balanced accuracy, calculated using leave-one-out for each rater. The classifier has a good performance compared to both experienced (A-D) and expert raters (E-G).

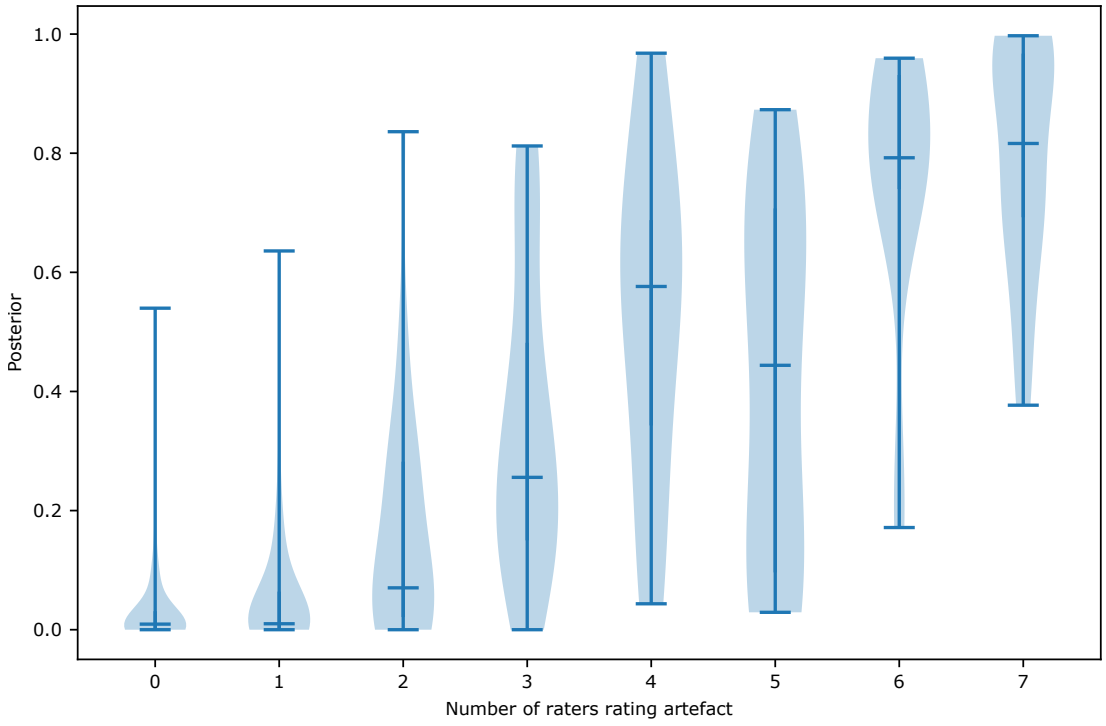


Figure 4.7: Violin plot showing the distribution of posterior probability against rater decisions. This shows the approximate shape of a logistic function. The posterior shown is the posterior probability of an epoch being artefact, as predicted by the algorithm. The number of raters rating artefact is from 0 (all raters agree epoch is clean) to 7 (all raters agree epoch is artefact). On each violin plot, the central line indicates the median of posteriors for that group, and the outer lines indicate the posteriors furthest from the median for that group.

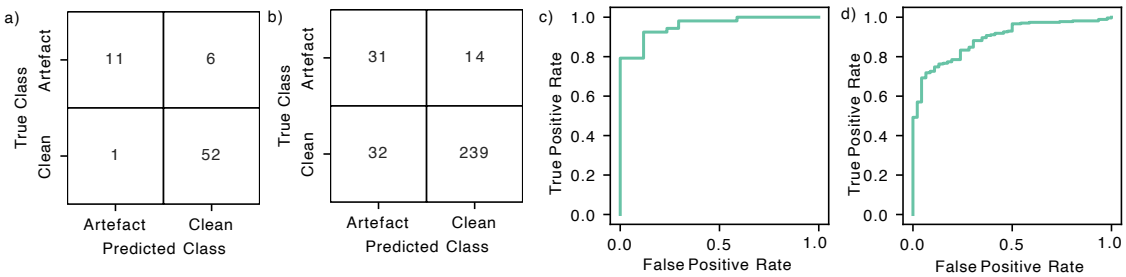


Figure 4.8: Testing the artefact detector in independent test sets. a) Confusion matrix showing performance of the artefact detector on novel data from the held-out test dataset: 0.81 balanced accuracy, 0.65 sensitivity, 0.98 specificity. b) Confusion matrix showing performance of the artefact detector on novel data from Dataset 2: 0.78 balanced accuracy, 0.67 sensitivity, 0.88 specificity. c) Receiver Operator Characteristic (ROC) curve, when tested on the held-out test dataset. Area under the curve: 0.96. d) Receiver Operator Characteristic (ROC) curve, when tested on novel data from Dataset 2. Area under the curve: 0.89.

4.4 Discussion

I have created an automated artefact detector for stimulus-evoked EEG data in infants and demonstrated that it performs well on single-subject data with low numbers of channels and short epochs. Accuracy is as good as manual review of data by individual raters, and it has the benefits of improved reproducibility, speed of analysis and could be used by non-expert users. Use of this method has benefits over those used by recent publications which present artefact identification algorithms (Debnath et al. 2020; Fló et al. 2022; Kumaravel et al. 2022), as it can be used on very short, single-channel epochs.

Current methods for artefact identification in neonatal EEG are not ideal. A number publications in the field involve the analysis of short epochs of low- or single-channel EEG data recorded from young infants, and most-commonly state that a manual review method is used to exclude artefactual data. Many publications do not report the methods used for artefact detection and exclusion, and often simply state the number of trials that were excluded (Fabrizi et al. 2011; Gursul et al. 2018; Norman et al. 2008; van der Vaart et al. 2019). Slater et al. (2010) used a 50 µV/50 ms (in >15 electrodes) ceiling, rejecting anything above that as movement artefact. Maimon et al. (2013) viewed 30-second increments of the average signal of Fp1, Fp2, Pz, C1, C2 and accepted anything "free of movements and electrical induction artefact". In Schmidt Mellado et al. (2022) a neurophysiologist excluded individual epochs "if the pre-stimulus baseline or the post-stimulus activity contained gross movement artefacts (e.g. signal amplitudes $> 800\mu V$ on any channel), or if the pre-stimulus baseline was unsettled (i.e. when stimulation was preceded by spontaneous bursting or other activity in the background or when EEG activity prior to stimulus was fluctuating$> 20\mu V$)". These examples highlight the range of different approaches that have been reported.

Neonatal event-related potentials can be used as outcome measures for clinical trials (Benoit et al. 2021; Cobo et al. 2022; Hartley et al. 2018; Slater et al. 2010b). Artefact can contaminate EEG recordings and thus the precision with which treatment effects can be identified is reduced, as it impacts the signal-to-noise ratio in the data.

By identifying and removing trials contaminated by artefact, the signal-to-noise ratio and the precision with which treatment effects can be identified can be improved. Importantly, by performing this identification using an automated classifier one can do this reproducibly across studies. This chapter shows that the proposed model generalises to other event-related potential datasets and so it has the potential to be used in future clinical trials that use event-related potential data.

The results show that the proposed method of automated artefact detection performs approximately as well as manual raters when applied to novel data. It performs well across a number of different types of stimuli, and across different neonatal age groups. Both posterior probability and accuracy are related to agreement between manual raters, suggesting that epochs which are clearly considered to be artefactual or not by manual raters are also highly likely to be correctly classified using the automated method. In addition, the automated method is performed very quickly (for a single sample, 477.55 ms $\pm$12.00 ms for feature calculation and 7.76 ms $\pm$0.06 ms for classification).

The performance of the artefact detection classifier was better for those epochs with high rater agreement (Figure 4.6d). This is as expected - those epochs where all raters agreed are likely to have overt artefacts more easily detectable, whereas epochs where only 4 out of 7 raters classed the epoch in one way (and the other 3 gave the opposite classification) are likely to have less obvious distinguishing features. Similarly, in random forest models, each individual decision tree produces a predicted class probability and the mean of all trees gives the posterior probability for a given class label.

One potential way of using this automated artefact detector is to use the automated method as a first review pass, and for the investigator to then review the segments of EEG data, together with the automated method's decisions and posterior probability, in order to confirm or dispute those decisions. Similar approaches are commonly used in the labelling of artefactual EEG components obtained through ICA (Pion-Tonachini, Kreutz-Delgado, and Makeig 2019). The investigator could then record those epochs for which the automated method's decision was disputed. This improves experimental reproducibility over existing methods. In addition,

compared to a single manual rater, this method should improve the accuracy of artefact identification and it could also be used as a training tool.

Interestingly, epochs labelled as artefact had on average a lower amplitude variance and measures correlated with it (such as sums of derivatives). This is unusual because EEG with artefact often has higher absolute signal magnitudes and variances (Delorme, Sejnowski, and Makeig 2007). As most epochs in this dataset contain event-related potentials, it is possible that some artefacts - such as electrode detachment - reduced the variance from event-related potentials. Alternatively Dataset 1 may contain relatively many low-variance artefacts such as skin-gel interface resistance or electrode detachment, and few high-variance artefacts such as blink, movement or ECG breakthrough. However, my work in this chapter did not classify artefacts by type.

Model performance - assessed using Leave-One-Subject-Out cross-validation - was not very sensitive to hyperparameter selection (Figure 4.1). Samples per leaf (Figure 4.1c) was the only hyperparameter whose optimisation visibly improved performance (although note that I have not calculated statistical significance for the effect of hyperparameter optimisation as part of this work). Lastly, permutation feature importance testing suggests that not all features add predictive value to the model (Figure 4.3) even though all features appeared to have some relation to the presence of artefact (Figure 4.5). Permutation feature importance can attach low importance values to features which are highly correlated with one another. This is because when a correlated feature is permuted, the model still has access to information about it through its correlated features. Most of the features used in the model were correlated with one another to some extent and so it is possible that permutation feature importance testing is under-estimating feature importance in correlated features.

The performance of a classifier trained using the popular vote of 3 'expert' raters (those with more than 4 years of experience) was not significantly different from that of a classifier using all 7 raters. Both had a balanced accuracy of 0.86, and using McNemar's Test $p = 0.65$. Training only using 'expert' decisions may have advantages - for example, the experience of expert raters may allow them to give more accurate decisions when compared to the group. On the other hand, training

in this way may have disadvantages - for example, the weighting of each observation (which was weighted by the fraction of raters who voted with the group) is less granular in a group of only 3 raters. The lack of a significant difference between classifiers may suggest that, if sufficient experienced raters are not available, then a smaller group of raters can be sufficient.

There are some limitations to this analysis. In particular, the relatively small size and imbalanced classes in the datasets used (with artefact being less prevalent than non-artefact observations) may limit the performance of the model. Imbalanced classes compound the effect of small dataset size because of the relatively small number of examples of the minority class. I reduced the impact of this by grouping and stratifying the training data in order to reduce over-fitting. Balanced accuracy on the held-out dataset (0.81) was similar to the balanced accuracy of the training dataset (0.86) which suggests that the impact of over-fitting was small. In the future, it is possible that performance may be improved with the use of additional training data. Importantly, this method also performs well on data recorded as part of a clinical trial and on individual measurements as would be recorded in a clinical setting and does not require subject-specific adaptation. This makes it potentially useful in clinical applications, where data from a single individual is considered.

Different applications have different requirements for artefact classification, and for some applications the probability threshold presented here is not appropriate. In some clinical uses a high specificity may be required in order to be as certain as possible that a clinical decision is not made based on data contaminated by artefact. On the other hand, in the context of a clinical trial it may be preferable to minimise the amount of data which is rejected in order to minimise the possibility of bias. For example, 68 epochs from the Petal clinical trial were incorrectly classified as artefact, which represented 22% of the included epochs (figure 4.8b). Had this approach been used in the trail this would have resulted in a large rate of data rejection. The original trial used a sample size selected for the observed variance in the data, and so this larger rate of data rejection would have unnecessarily reduced the power of the trial. In order to adjust the classifier for different priorities, a user can adjust the probability threshold at which an epoch is classified as artefact. The

effect of this is shown in figures 4.8c and 4.8d. In this way a sensitivity-specificity balance can be chosen for a particular application.

I designed this classifier specifically for use in single-channel EEG data. Hypothetically the classifier can be applied to any number of channels (it is not channel-specific), but this has not been validated here. In particular, manual rating of artefact in multi-channel EEG often relies on inter-channel information (e.g. in identifying ECG break through in the recordings) and many existing automated methods use channel covariance information. This limitation is not a design flaw, but a result of the decision to focus on the research gap which is artefact detection for single-channel data. The method which I present here by design does not consider inter-channel data.

In conclusion, I created and tested an automated method of detecting artefact in individual, single-channel epochs of infant EEG. It fills a niche in the automated identification of artefact in low-channel event-related potential analysis. This method has the benefits of reproducibility and speed of analysis when compared to the manual methods that are commonly used.

5

A clinically usable measure of infant nociception

Contents

5.1 Introduction 74
5.1.1 Objective 74
5.1.2 Biomarkers of neonatal noxious response 75
5.1.3 A diagnostic biomarker 77
5.1.4 A monitoring biomarker 78
5.1.5 Validation 78
5.1.6 Chapter outline 81
5.2 Methods 83
5.2.1 Datasets and study design 83
5.2.2 Response magnitude estimate 85
5.2.3 Experimental stimuli 87
5.2.4 Fit normalisation 87
5.2.5 Baseline standardisation 88
5.2.6 Heartrate modulation 89
5.2.7 Evaluation metrics 92
5.3 Results 94
5.3.1 Response magnitude estimate 94
5.3.2 Experimental stimuli 94
5.3.3 Fit normalisation 96
5.3.4 Baseline standardisation 97
5.3.5 Heartrate modulation 99
5.3.6 Evaluation using an independent validation dataset 99
5.4 Discussion 103

Related publication: Slater et al. 2023
Part of this chapter was published in the above patent, publication number WO2023084208A1. I am an inventor of the patent, and the content of this chapter is entirely my own.

5.1 Introduction

5.1.1 Objective

In this chapter I will introduce a scale of infant noxious response, based on ERP response magnitude estimation. I propose four modifications to the scale, and assess each one using known-groups validity. I compare each modification to the unmodified scale.

The objective is to assess the known-groups validity of these noxious response scales which are diagnostic and monitoring biomarkers. Two separate pairs of known groups are used:

- Heel lance (an acute somatic nociceptive skin-breaking clinical procedure) against control heel lance (an acute somatic non-nociceptive non-skin-breaking experimental procedure).
- Experimental stimuli (an acute somatic nociceptive non-skin-breaking experimental procedure) against background (no procedure).

Research question: For an acute somatic clinical procedure, a) is the difference (noxious minus non-noxious) in individual modified response magnitude estimates significantly greater than the individual unmodified response magnitude estimates?, and b) when discriminating between a noxious and non-noxious acute somatic clinical procedure, is the area under the Receiver Operating Characteristic curve for the modified response magnitude estimates significantly greater than for the unmodified estimates?

5.1.2 Biomarkers of neonatal noxious response

The previous chapters of this thesis have primarily dealt with developing methods for use in clinical research. This is a major component of the effort to prevent unnecessary pain in neonates, as evaluation of analgesics in the neonatal population requires a suitable biomarker. Biomarkers for this purpose tend to be response biomarkers (see Section 2.2.4), used to show whether an effect has occurred in individuals within a group treated with analgesics. Using a method within this context requires certain assumptions. Each infant will be studied as part of a group of individuals, and the group average effect is of interest rather than the effect on individuals. This is because evaluating a drug for a population requires that results can be generalised from the study group to the wider population, rather than being specific to any particular individual (Bland 2015; Cook 2023). The majority of research into neuroimaging biomarkers for infant nociception has addressed this context, and has developed methods for clinical research.

A separate context for the use of neuroimaging biomarkers is clinical use. However, the clinical use of these methods has not yet been adequately addressed in the literature: methods which are suitable for clinical use do not yet exist. Cowen et al. (2015) explain that "at present, these techniques remain within the research setting, requiring careful experimental protocols, a sophisticated extraction process and high computational complexity. They are, therefore, beyond the realms of clinical application, but may provide possible avenues for development in the future" (Cowen et al. 2015).

This context is important for the ultimate goal of preventing unnecessary pain in infants. Infants who are routinely monitored in an inpatient setting tend to be sick, by the nature of the setting. Sick or premature infants can be unable to mount a normal behavioural and physiological response to painful stimuli, due to external factors such as ventilation and medication or due to internal factors such as lack of energy or altered development (Barr 1998; Green et al. 2019; Sellam et al. 2011).

A method for identifying nociception in infants in a clinical setting would be useful as it would allow us to identify whether an individual infant was experiencing pain from a particular stimulus. This would primarily help with two use-cases.

1. **Diagnostic biomarker.** In this use-case EEG is time-locked to a noxious clinical event, such as an immunisation or a heel lance for blood tests. Responses to noxious stimulation vary both between infants and in individuals over time. A method for identifying nociception in this context would allow clinical staff to identify procedures which are currently causing a nociceptive response in an individual infant.

2. **Monitoring biomarker.** In this use-case a series of non-clinical, non-skin-breaking, noxious stimuli is applied to an infant by clinical staff. By measuring the magnitude of the infant's cortical nociceptive response, this method allows staff to monitor an infant's susceptibility to nociceptive response. This allows non-time-sensitive clinical procedures to be performed at an appropriate time, in order to minimise an infant's discomfort.

It has been suggested that EEG be used for routine monitoring of hospitalised neonates (Vanhatalo and Lauronen 2006), and that this may provide a suitable quantitative method for identifying nociception in infants (Hartley and Slater 2014). The magnitude of the ERP has previously been shown to relate to noxious response (Hartley et al. 2015, 2017). Similarly, heartrate change has previously been shown to improve classification of noxious from non-noxious stimuli (van der Vaart et al. 2019).

When it comes to clinical instrument development, the framework for evaluating the measurement properties of an instrument laid out by COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) is the current gold standard (de Vet et al. 2011; Mokkink et al. 2010). To establish the validity of this new measurement approach, I have adopted the COSMIN terminology. A popular alternative framework is that of the United States Federal Drug Administration (FDA) (FDA 2009). It is worth noting the FDA terminology maps directly onto that of COSMIN for validity. However, the FDA framework has been developed with

patient-reported outcome (PRO) measures in mind, while the specific EEG metric under investigation is a biomarker (FDA-NIH Biomarker Working Group 2021).

Validity is defined by the COSMIN panel as 'the degree to which an instrument truly measures the construct(s) it purports to measure' (de Vet et al. 2011; Mokkink et al. 2010). There are several subgroups of test validity - see Figure 5.1 below. The type of validity most appropriate for the current purpose is a subtype of construct validity called known-groups validity. Known-groups validity is demonstrated when a test or questionnaire can discriminate between two groups known to differ on the variable of interest (Davidson 2014).

5.1.3 A diagnostic biomarker

Here I focus on two aspects of a measure's usability in clinical settings: reliability and simplicity. In order to improve the usability of the measure in a clinical setting we wish to improve the reliability of the measure. In this context I use the COSMIN definition of reliability: the proportion of measurement variance which is due to 'true' differences between measured values (Mokkink et al. 2023, 2010). In a group analysis such as in a clinical trial, any spurious responses make up a relatively small proportion of total responses - assuming a normal distribution and a suitable sample size - and so are unlikely to unreasonably skew the data. In contrast, in clinical care a single measurement could be all that can be acquired for an indivial patient: a spurious reading could at best reduce the confidence of clinical staff in the biomarker, and at worst has the potential to inappropriately change the course of clinical management for the patient. In order to reduce the likelihood of these adverse events, a good reliability is desirable in a clinical setting. In addition, in a research setting multiple variables can be calculated and appropriate statistical treatment applied. In contrast, a clinical setting is time-constrained and requires information to be presented to staff in a simple and understandable way in order to reduce cognitive overhead for decision making. Complexity can lead to mistakes in clinical care due to human error, and this is to be avoided by design if possible. In order to reduce the potential for human error arising from measurement complexity, a small number of displayed measures (ideally a single measure) is desirable in clinical settings.

5.1.4 A monitoring biomarker

A diagnostic biomarker allows identification of noxious response to events, and heel lances are a type of noxious event which is common clinically. A monitoring biomarker introduces an additional complication: response magnitude estimation requires a time-locked stimulus to measure the response to. Using response magnitude estimation as a monitoring biomarker therefore requires a stimulus to be applied whenever monitoring is needed. There are no clinical procedures for inpatient neonates which are known to be noxious and can be assumed to be performed frequently and regularly. It is not appropriate to perform heel lances when not clinically necessary, so an alternative is required. One option for experimental stimuli which has been evaluated previously is a low intensity experimental noxious stimulus.

Regardless of how well-adjusted the data are, there will be unavoidable measurement error: for example, noise increasing or decreasing EEG deflection, or an unrelated event increasing the heartrate. One way of reducing the impact of this measurement noise is to perform multiple measurements on an individual infant, and this is feasible with low intensity stimuli. Averaging a number of measurements may reduce the impact of this error.

The time of clinical staff is valuable. Where it is possible to average over multiple measurements, there is a trade-off between measurement accuracy and time taken to take the measurement. This trade-off is inherently context-specific and somewhat subjective, and in order to better judge this I estimated measurement error for different numbers of measurements.

5.1.5 Validation

We use the COSMIN framework (Mokkink et al. 2010) for definitions regarding validation of clinical instrument measurements, as described in 2.2.5. This framework is separate to the validity frameworks used in other fields, such as in machine learning.

The COSMIN framework delineates 3 broad types of validity: content validity, criterion validity, and construct validity (Figure 5.1). Content validity (the degree to which the content of an instrument is an adequate reflection of the construct

which it purports to measure) is primarily qualitative. It is not established from quantitative analysis of the instrument but is established using, for example, panel reviews of experts. This type of qualitative validation is outside the scope of this chapter. Criterion validity (the degree to which the scores of an instrument are an adequate reflection of a 'gold standard') is measured quantitatively against an existing gold standard. Currently, no gold standard for pain assessment exists (Baxter et al. 2024; Glenzel et al. 2023; Llerena et al. 2023) and so we cannot validate this biomarker using criterion validity. Construct validity (the degree to which instrument scores are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured) can be quantitatively assessed using a hypothesis-driven approach, and so I will validate the biomarker using this type of validity.

The COSMIN framework split construct validity split into 3 subgroups: structural validity, cross-cultural validity, and hypothesis testing validity (Figure 5.1). Structural validity (the degree to which instrument scores are an adequate reflection of the dimensionality of the construct to be measured) is only relevant to multi-dimensional instruments and so is not relevant for this biomarker which has only one dimension. Cross-cultural validity (the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version) is also irrelevant as there is no language component to be translated and there are no aspects which have been culturally adapted. Hypothesis testing validity (which has a COSMIN definition identical to that of construct validity) can be measured for this instrument, as explained above.

The COSMIN framework gives a non-exhaustive list of three possible types of validation within hypothesis testing validity: convergent validity, discriminant validity, and known groups validity (Figure 5.1). Convergent validity (correlation with similar constructs) could be measured for this biomarker. However, I will not do so here. As previously explained, there is no gold standard for infant pain assessment. Further, different instruments are based on different mechanisms (e.g. behaviour, heart rate or EEG) and so different instruments need not agree in order for them to be related to some aspect of the pain experience. For this reason I will

not carry out convergent validation in this chapter. Discriminant validity (lack of correlation with unrelated constructs) could also be measured for this biomarker. This would require that an unrelated construct was also measured at the same time as the stimulus response. This was not the case in the data used for this chapter, and collecting new clinical data is beyond the scope of the chapter. For this reason I will not carry out discriminant validation in this chapter. Lastly, known groups validity (discrimination between groups whose relevant characteristics are known) can be measured for this biomarker. We can expect a group of responses to noxious stimuli to be different to a group of responses to non-noxious stimuli. I will validate the biomarker using known-groups validity, by testing its ability to differentiate between noxious and non-noxious stimuli.

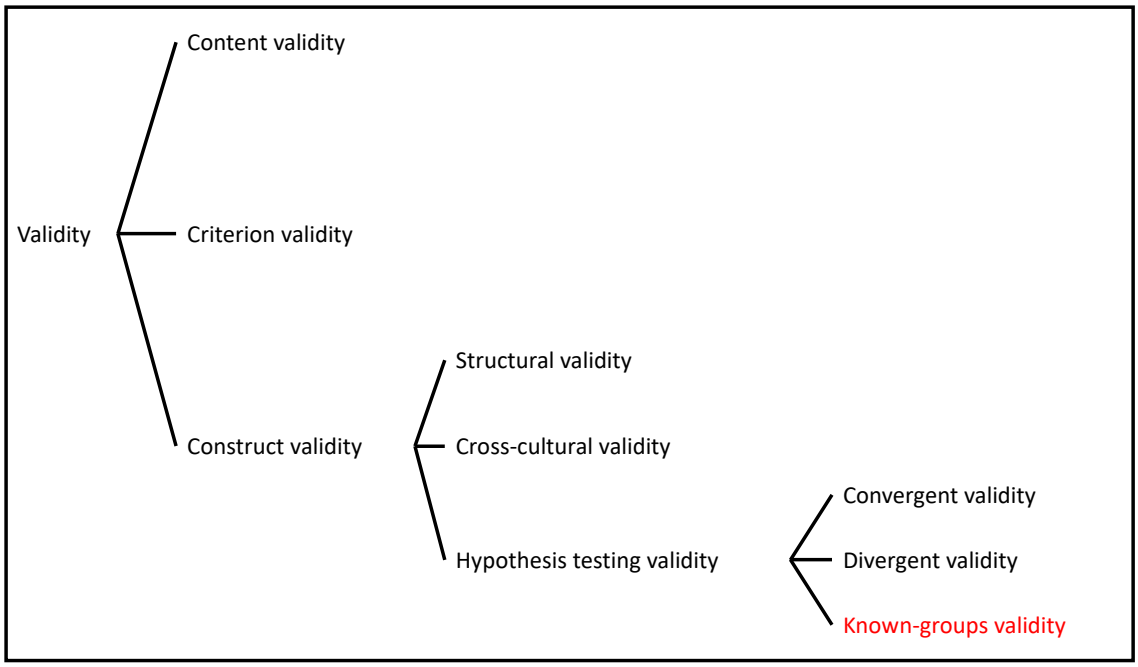


Figure 5.1: Types of validity in the COSMIN framework. Known-groups validity, the only type of validity which is feasible to test for this biomarker, is highlighted with red text. Image adapted with permission from Baxter et al. 2024.

I use heel lances for validation of a diagnostic biomarker because these are skin-breaking procedures and so are unambiguously related to pain as they are associated with actual tissue damage (Raja et al. 2020). Heel lances are recommended as an acute pain model for clinical trials by the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION)

Pediatric Pain Research Consortium consensus meeting (Walco et al. 2018). Control heel lances contain the same acoustic and tactile stimulus properties as a heel lance, without the noxious component (as there is no skin-breaking lance). For a diagnostic biomarker a clinically-required heel lance can be used, which ensures that no skin-breaking procedures are carried out unless already clinically necessary.

I use experimental noxious stimuli for validation of a monitoring biomarker because the stimulus must be applied any time that a monitoring measurement is desired, and application of a skin-breaking stimulus such as a heel lance would be unethical unless clinically necessary. An experimental noxious stimulus is an acute somatic nociceptive non-skin-breaking experimental procedure. There is no procedure available which contains the same non-noxious sensory qualities as an experimental noxious stimulus, and so a background period (i.e. a period where no stimuli are applied by the researcher) is used as a control for this procedure.

5.1.6 Chapter outline

This chapter aims to develop a scale representing an infant's response to noxious stimulus. The scale is based on the estimated response magnitude calculated from the event-related potential using the method outlined by Hartley et al. (2017). The main benefit of a biologically-grounded scale compared to more data-driven alternatives is that a scale such as this is likely to generalise to new stimulus types.

The estimated response magnitude is modified in the following ways in order to create a new scale of noxious response. This scale is intended to be simple (represented by a single, positive real number) and reliable (a low proportion of measurement error to true differences in measure).

- **Baseline standardisation:** When the waveform from Hartley et al. (2017) is applied to an individual, the estimated response magnitude is not directly comparable between individuals. This is primarily because the magnitude of electrophysiological activity recorded using scalp electrodes differs between individuals, due to confounding variables such as electrode placement, skin conductance, resistance at the skin-electrode interface, head and cortical

topography, and individual excitability. EEG from a stimulus-free period can provide a baseline with which to compare stimulus responses. In order to do this, response magnitudes are estimated from a stimulus-free baseline period in the 10 seconds immediately preceding the first stimulus, in the same way as for stimuli. This gives us a distribution of estimated response magnitude values, for an individual infant, when there is no stimulus. A stimulus response can then be standardised to an individual's baseline by calculating the response's z-score using the baseline distribution (Equation 5.5).

- **Fit normalisation:** Where the neurological signal in the response to stimulus is much smaller than noise, it is possible for the template to be fitted to a spurious signal. This can cause artefactually high or low response magnitude estimates. In order to mitigate this we can adjust a magnitude estimate for the template fit by normalising it to the Pearson correlation coefficient between the corresponding section of EEG signal and the scaled noxious template (Equation 5.4).

- **Heartrate modulation:** A method of increasing our confidence in the measure which we produce is to introduce a second measurement which we believe is related to the noxious response but which is a separate modality from the cortical activity which we are measuring. By using multimodal data in this way we hope to improve measurement reliability. Heart rate is one possible measure for this. Recent work by van der Vaart et al. (2019) has shown that using multiple modalities of physiological measurement can improve the ability to discriminate between noxious and non-noxious stimuli. In this study, facial expression, EMG (reflex withdrawal), EEG, ECG (heartrate) and SpO2 were measured. Of these, both heartrate and SpO2 are already routinely monitored in some inpatient neonatal settings and heartrate (but not SpO2) was found to provide good discrimination between noxious and non-noxious events. This chapter uses the feature identified by van der Vaart et al. (2019) to modulate the estimated response magnitude.

At each step I evaluate the known-groups validity of the measure on two sets of known groups: clinical heel lance vs. control heel lance, and experimental noxious stimulus vs. stimulus-free background.

I evaluate the measures using an independent validation dataset. This ensures that the results are generalisable to this separate dataset. Using a separate dataset for testing eliminates the bias which might arise from evaluating the scales using the same dataset as I developed them with. I evaluate the scales with comparison to the unmodified scale, using the metrics described in *Evaluation Metrics*.

In summary, I investigate how to improve ERP response magnitude estimation for use in clinical care as a diagnostic biomarker (using response to heel lance as an example) or as a monitoring biomarker (using response to mild experimental stimulus as an example). This chapter investigates the known-groups validity of four possible methods for improving this diagnostic and monitoring biomarker.

5.2 Methods

5.2.1 Datasets and study design

This chapter uses three datasets. I use Datasets 3 & 4, described in Chapter 3, for development of the different noxious response scales. Dataset 3 contains a pair of single, clinical noxious stimuli for each participant. Dataset 4 contains multiple non-clinical noxious stimuli for each participant. I use Dataset 2 as a validation set. Dataset 2 contains a pair of single, clinical noxious stimuli for each participant. This dataset is the same as the validation dataset used in Chapter 4. A summary of the datasets which I use in this chapter is shown in Table 5.1.

Datasets 3 and 4 are used to develop and test each modification of the noxious response scale. Evaluating these methods using the same dataset carries the risk of over-fitting. The known-group validity of each noxious response scale can be evaluated in an independent, validation dataset. Validation with an independent dataset ensures that the scales are generalisable beyond the original datasets. I validated the scales using Dataset 2. This dataset is from a previously conducted

multi-site clinical trial, and includes paired epochs from clinically-required heel lance and control heel lance. The dataset was previously used in Chapter 4.

I will briefly describe the new datasets. Dataset 3 consists of epochs around paired noxious stimuli (heel lance stimuli) and non-noxious controls (control heel lance stimuli). I chose heel lances for initial analysis as they are a common clinical stimulus, well-characterised, easy to time-lock, and provide an acute noxious-evoked response that can be discriminated from non-noxious events. Control heel lances provide a non-noxious stimulus which still applies the vibrotactile and auditory stimulus as the heel lance itself. Dataset 4 consists of epochs around noxious stimuli (experimental noxious stimuli) and non-noxious controls (manually-labelled "background" epochs). Experimental noxious stimuli have the benefit of being able to be applied multiple times to an infant within a single setting; this allows averaging over multiple stimuli. Background epochs provide a non-noxious control for experimental stimuli: each epoch was manually labelled during original data collection as a period without external stimulation.

The following inclusion criteria were used. These criteria are deliberately broad, to best mimic clinically-gathered data and to ensure a robust measure.

- GA>36 weeks at birth
- EEG and ECG collected
- Data is free from artefact, as identified by a clinician experienced in identifying artefact in research settings (checked at stimulus ± 2 seconds)

I band-pass filtered ECG data to 12-40Hz and separated it into 20-second long epochs, 10 seconds before and 10 seconds after each event. This epoch duration is in line with similar, previous work (van der Vaart et al. 2019). I performed R-peak detection using the prominence-based algorithm of MatLab's `findpeaks` command. 3 second average R-R interval was calculated every 1 second and the inverse of the R-R interval gave heartrate.

I carried out EEG data pre-processing using the pipeline described in Appendix A. Data was band-pass filtered to 0.5-30Hz (Luck 2014). I then epoched the EEG

	Clinical stimuli	Experimental stimuli
Development	Dataset 3	Dataset 4
Validation	Dataset 2	N/A

Table 5.1: Summary of the datasets used in this chapter. Clinical stimuli consist of heel lance (an acute somatic nociceptive skin-breaking clinical procedure) and control heel lance (an acute somatic non-nociceptive non-skin-breaking experimental procedure). There is one pair of clinical stimuli for each participant. Non-clinical stimuli consist of experimental stimuli (an acute somatic nociceptive non-skin-breaking experimental procedure) and background (no procedure). There are multiple non-clinical stimuli for each participant.

into 20-second long epochs, 10 seconds before and 10 seconds after each event. This epoch length maintained consistency with ECG epochs.

5.2.2 Response magnitude estimate

The work of Hartley et al. (2017) shows a template of noxious-evoked brain activity, produced as described in 2.2.3. This method consists of two steps. In the first step the template is first Woody filtered. Woody filtering (Woody 1967) calculates cross-correlation between signal and template in order to identify the latency with the maximum cross-correlation. The Hartley method limits the possible latencies to a small range in order to ensure that the latency of the presumed event-related potential is biologically plausible. In the original publication this restriction is ±50ms, although in this chapter I use ±100ms in line with more recent work (Cobo et al. 2021; Hauck et al. 2024; Schmidt Mellado et al. 2022). The second step is Singular Value Decomposition, which fits the template onto the data at the chosen latency. Singular Value Decomposition of template X_0 is described by Equation 5.1. The weights provided by applying the Singular Value Decomposition to new data represent the magnitude estimates of the event-related response.

$$X_0 = U_0 \Gamma_0 V_0^T \tag{5.1}$$

This results in a single real number for each stimulus, which quantifies the magnitude of the scaled template that best fits the signal. Figure 5.2 shows approximate distributions of these response magnitude estimates, for noxious heel lances and for non-noxious, tactile control heel lances.

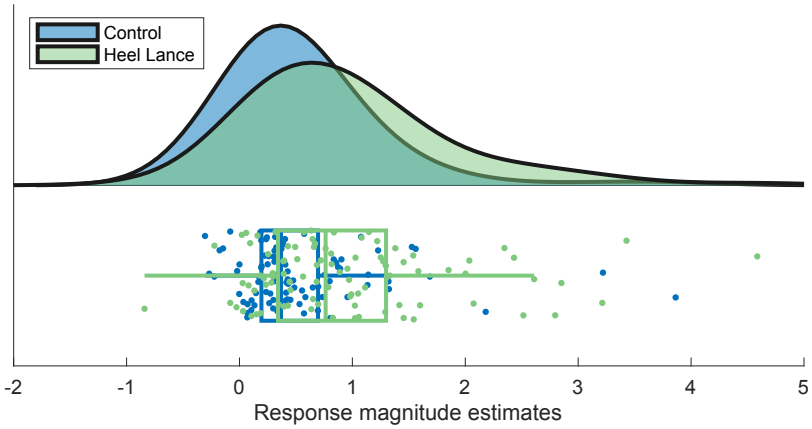


Figure 5.2: Raincloud plot showing response magnitude distributions in Dataset 3. Noxious heel lances and tactile controls are shown here. Box plots below the distributions show median and interquartile range. Note that this is a different dataset to those shown in Figures A.5 and A.6.

I applied the method of Hartley et al. (2017) to the dataset in order to establish an unmodified 'control' method to which we can compare our modifications. Using a paired t-test we check whether the estimated response magnitude is higher, on average for the dataset, after noxious stimuli than after non-noxious control stimuli. I selected a template coefficient value to be a cut-off such that the template coefficients above this threshold predict a noxious event, and coefficients below predict a non-noxious event. By varying the threshold of the template coefficient and taking the True Positive Rate (sensitivity of predictor to nociception) and False Positive Rate (inverse of specificity of predictor to nociception) for each threshold, a Receiver Operating Characteristic curve was plotted to assess the template coefficients diagnostic value.

The 'unmodified' response magnitude estimate is the current state-of-the-art. This is the metric which I will build upon and which I will compare to. Other methods are available. For example: a classification model would allow us to optimise the known-groups validation; alternatively a regression machine learning model would allow us to predict a pain 'level' if we were to have some gold-standard method for identifying a ground truth of pain for different types of stimulus. The main benefit of making a biologically-grounded scale, compared to data-driven methods such as the ones just described, is that it is likely to later generalise to other types of stimulus which are not available for testing.

5.2.3 Experimental stimuli

Experimental stimuli such as the ones used here apply lower force than clinical heel lances, and have been shown to elicit a noxious response (Hartley et al. 2015), but not a behavioural one. This response is graded by stimulus intensity, with experimental stimuli giving smaller magnitude responses. As such it may be more difficult to differentiate between noxious and non-noxious events. With experimental stimuli, however, we have an opportunity to collect data from multiple stimuli from a single infant. For a single infant, I average the response magnitude from multiple identical stimuli. The subject-level measure is then this average.

In order to investigate the effect of this averaging, we use the data described previously. For each measurement on each subject, we will have a long-run average magnitude (Equation 5.2). This is the average (μ) of the response magnitude estimates (T) over the full number of stimuli (m) presented to that infant.

$$\mu[T_i]_1^{i=m} = \frac{\Sigma T_i}{m} \tag{5.2}$$

This number of responses takes an unnecessarily long period of time to measure. I calculate measurement error of the cumulative mean by averaging the measurements so far, normalising to the long-run average, and normalising that to the standard deviation of all measurements. The calculation used is shown below, where m is the total number of recorded template coefficients for a given infant, and n is the number of coefficients to find the average of.

$$T_{error}|_{n=1:m} = \frac{abs(\mu[T_i]_1^{i=n} - \mu[T_i]_1^{i=m})}{\sigma[T_i]_1^{i=m}} \tag{5.3}$$

For this analysis, subjects were not subject to a sham event (i.e. the control heel lance stimuli) so I use background data as the non-noxious control.

5.2.4 Fit normalisation

The estimate of the response magnitude is the scaling factor of the template which best fits the data, but a good fit does not necessarily follow from a large scaling factor. As an example of how this can impact results, Figure 5.3 shows EEG traces

in four epochs containing a stimulus at t=0. The Woody-filtered and scaled template is superimposed onto the EEG signal. Figures 5.3a and 5.3b show epochs with low estimated response magnitudes, with low Pearson correlation (5.3a) and high Pearson correlation (5.3b) respectively. In the same vein, Figures 5.3c and 5.3d show epochs with high estimated response magnitudes, with low Pearson correlation (5.3c) and high Pearson correlation (5.3d).

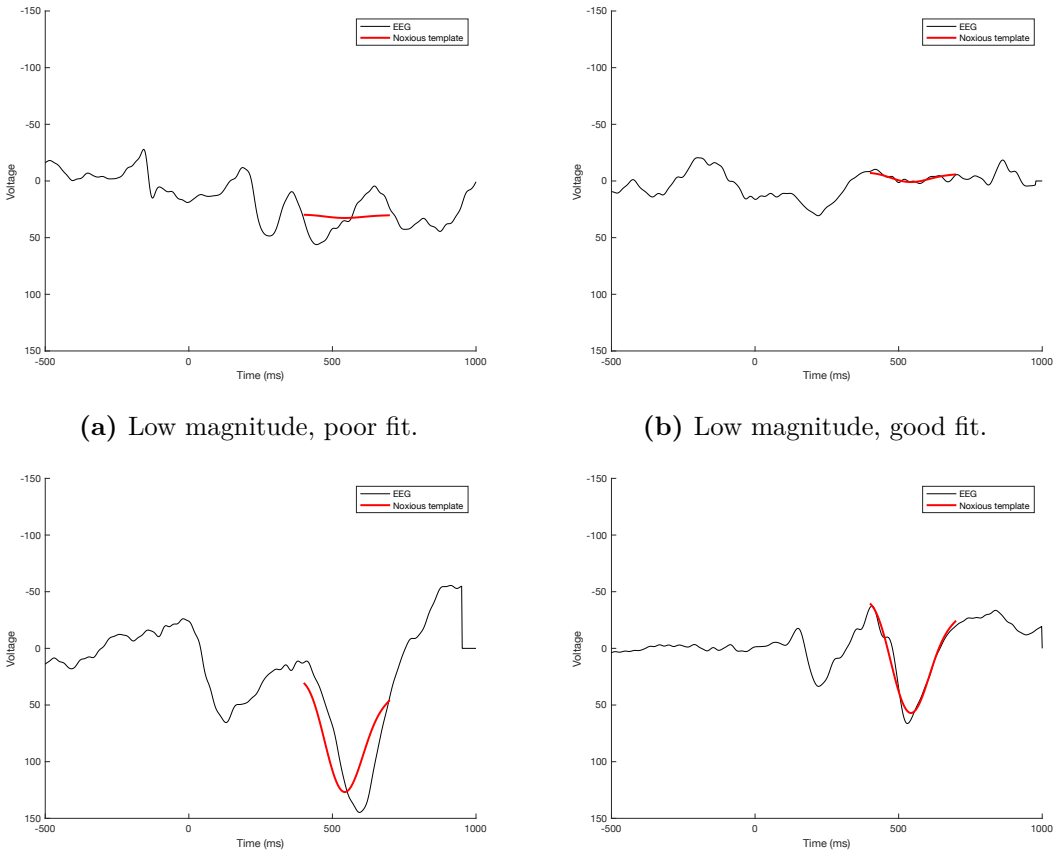
EEG data is typically visually cleaned, but an automated method of removing poor fits reduces the time required to take a measurement in a clinical setting. An automated method was developed in Chapter 4, and this provides a fast and reproducible decision on which epochs to remove because of artefact. However, we do not always want a binary decision. To adjust for our confidence in an estimated response magnitude, we normalise the template coefficient by the Pearson correlation coefficient at best fit, where normalised template coefficient is given in Equation 5.4 where T is the response magnitude estimate and p is the Pearson correlation coefficient. *template* and EEG represent the vectors of the scaled template signal and the EEG signal, respectively.

$$T_{norm} = T \cdot \rho(template, EEG) \tag{5.4}$$

5.2.5 Baseline standardisation

One of the drawbacks of using the magnitude of a fitted signal in a score such as this is that it is possible that the magnitude of these responses could be dependent on the magnitude of the signal amplitude variance. This is because increased magnitude of amplitude variations is more likely to produce higher template coefficients by fitting to higher amplitudes of scaled template.

We can apply magnitude response estimation at arbitrary points where we expect there to be no noxious response - for example during the baseline prior to noxious stimulus. The average template coefficient at this point should be very low, but not zero as any varying signal (such as background EEG or non-neural noise) will have a non-zero correlation to a scaled template. This is because Woody filtering optimises to the maximum positive fit, which skews the average towards positive values. We



(a) Low magnitude, poor fit.

(b) Low magnitude, good fit.

(c) High magnitude, poor fit.

(d) High magnitude, good fit.

Figure 5.3: Examples of response magnitude estimates with different Pearson's correlations (fits). Each subplot shows a noxious activity template (red) superimposed on EEG following noxious heel lance (a and d) or tactile control (b and c).

tested a method of standardisation to a baseline that was the ten seconds pre-stimulus, resulting in standardised magnitude response estimate T_{std} shown in Equation 5.5.

$$T_{std|t=0} = \frac{T|_{t=0} - \mu[T|_{t=-10:0s}]}{\sigma[T|_{t=-10:0s}]} \tag{5.5}$$

This is chosen over the simpler method of normalising to baseline variance ($T_{std|t=0} = \frac{T|_{t=0}}{var[HR|_{t=-10:0s}]}$) as it is more directly linked to the metric of interest.

5.2.6 Heartrate modulation

One method of increasing our confidence in the data is to introduce a separate,noxious-related modality. This allows us to rely less on a single measure, as we can adjust our confidence in the EEG-based results based on this other associated metric. In

the work of van der Vaart et al. (2019) a heartrate statistic, as below, was found to differentiate well between noxious and non-noxious stimuli. This statistic finds the maximum heartrate in the 15 seconds following stimulus and subtracting the pre-stimulus mean heartrate. In many cases epochs of ±15 seconds overlap with previous or later epochs. To avoid this here we use 10 seconds before and after the stimulus.

$$HR_{stat} = max(HR|_{t=0:10sec}) - \mu(HR|_{t=-10:0sec}) \tag{5.6}$$

All epochs in the datasets used here have associated ECG, recorded synchronously with EEG using the infants' cot-side vital signs monitors. ECG is high-pass filtered at 12 Hz and low-pass filtered at 40 Hz. I calculate heart rate by finding local ECG maxima (matlab `findpeaks`) that are at least 1.5 standard deviations from the mean of the filtered ECG signal. Maxima less than 350 ms from a previous maximum are discarded, as heart rates above 170 beats per minute would not be expected in this population. The `findpeaks` method (Sadek et al. 2017) is used because the code is open-source and, when a reasonable minimum peak distance (i.e. maximum heart rate) is used, it has acceptable accuracy compared to other methods when identifying QRS complexes from ECG and ballistocardiograms (Shimauchi et al. 2021; Suliman et al. 2019).

From the work of van der Vaart et al. (2019) we can expect that EEG response magnitudes should generally be correlated with standardised heartrate changes. In order to estimate the expected relationship between the EEG response metric (Equation 5.1) and the heart rate response metric (Equation 5.6), I fit a linear model to heel lance epochs. This gives a line with gradient 0.0096 and intercept 0.57. Conceptuallyhis means that the EEG response is slightly related to the heartrate statistic ($R^2 = 0.025$, $p = 0.033$), and from this relation - for a given heartrate statistic, HR - you would expect the EEG response magnitude to be $0.0096HR+0.57$.

In order to create a scale from these two measures, I modulate the EEG response metric by this expected response (shown in Figure 5.4). This modulation is shown in Equation 5.7 where y is the EEG response magnitude, x is the heartrate statistic (Equation 5.6), p is the proportional contribution of the EEG response magnitude

(between 0 and 1) and f is the function itself. This can be visualised using a plane with estimated EEG response magnitude (y axis) and calculated heart rate statistic (x axis). An illustrative example is shown in Figure 5.4.

$$f(x, y) = py + (1 - p)(0.0096x + 0.57) \tag{5.7}$$

The EEG and heartrate metrics can contribute to the modulated result in different amounts (represented by p in Equation 5.7). I optimised the value of p for the largest t-statistic in a paired t-test comparing heel lance to tactile control, as this maximises the separation between the two groups. Of 11 values of p (from 0 to 1 in increments of 0.1), the optimal value was $p = 0.3$ (i.e. EEG contributes 30% and heartrate contributes 70%). The optimisation is shown in Figure 5.5.

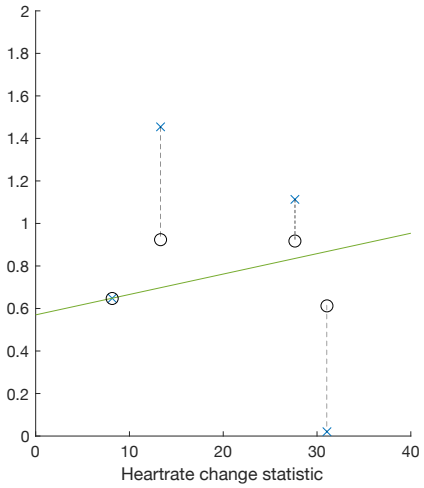


Figure 5.4: Illustrative example of heart rate modulation for four exemplar heel lance epochs. Crosses indicate unmodified EEG response magnitude estimates, with circles indicating the estimate after the appropriate heartrate modulation. Estimates are modified to be closer to the value which would be predicted from the heartrate.

I will not evaluate the performance of the heartrate statistic alone. This is because cortical responses are specific to noxious stimuli (Fabrizi et al. 2011; Hartley et al. 2017) whereas heartrate changes with other stimuli such as auditory, tactile and visual and so it is not specific to noxious response (Pomerleau-Malcuit and Clifton 1973; Von Bargen 1983). A measure based solely on heart rate would not be specific to noxious stimuli. For more information regarding the benefits of cortical measures, see Section 2.3.3.

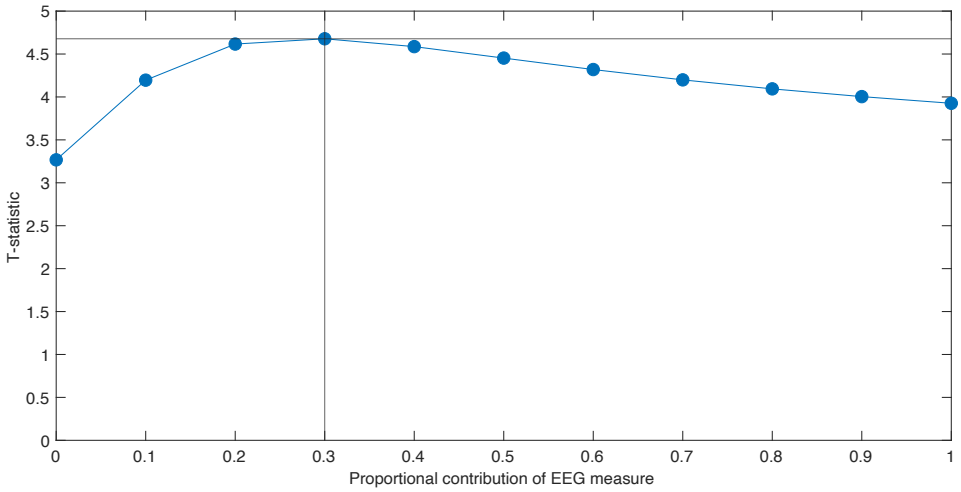


Figure 5.5: Results of optimisation of the parameter p in proportional modulation, where p is the proportional contribution of the EEG metric (Equation 5.7). $p = 0.3$ produces the optimal t-test statistic (4.68).

5.2.7 Evaluation metrics

The goal of this exercise is to assess the known-groups validity of a noxious response scale. As described in the Objectives section, I use two separate pairs of known groups. These groups are as follows.

- Heel lance (an acute somatic nociceptive skin-breaking clinical procedure) against control heel lance (an acute somatic non-nociceptive non-skin-breaking experimental procedure).
- Experimental stimuli (an acute somatic nociceptive non-skin-breaking experimental procedure) against background (no procedure).

For this I use the paired t-test. The paired t test is shown in equation 5.8, where β is a vector of the pairwise differences between values x and y, SE indicates the standard error ($\frac{\sigma}{\sqrt{n}}$), and t is the t statistic. The paired t test calculates a t statistic, which is the ratio of the difference between two parameters (here this is estimated magnitude) and their standard errors.

$$t = \frac{\bar{\beta}}{SE(\beta)};\ \beta_i = y_i - x_i \tag{5.8}$$

For each scale in this chapter, I calculate two summary statistics for each dataset. Each is evaluated by comparing metrics for noxious and non-noxious stimuli.

1. **Paired t-test t-statistic.** Here the t test uses the difference between noxious and non-noxious metrics, for a single scale. For this t test, x is the response magnitude score after a non-noxious stimulus ($x = T_{nonnoxious}$) and y is the response magnitude score after a noxious stimulus ($y = T_{noxious}$). The t statistic represents the difference between the noxious and non-noxious metrics on this scale.

2. **Area Under the Receiver Operating Characteristic curve (AUROC).** When an ROC is calculated using a parameter (estimated magnitude), the AUROC is the fraction of the total area which is under that curve. Here it represents the differentiability of the noxious from non-noxious groups. Its lower bound is 0.5 (assuming that it is better than chance).

As described in the *Objectives* section, the goal is to assess known-groups validity, i.e. the degree to which a scale discriminates between two groups known to differ. The research question, as described earlier, is as follows. *For an acute somatic clinical procedure, a) is the difference (noxious minus non-noxious) in individual modified response magnitude estimates significantly greater than the individual unmodified response magnitude estimates?, and b) when discriminating between a noxious and non-noxious acute somatic clinical procedure, is the area under the Receiver Operating Characteristic curve for the modified response magnitude estimates significantly greater than for the unmodified estimates?*

In order to answer part (a) of the research question, I compare noxious to non-noxious differences between scales. For this the t test uses the difference between the noxious-to-non-noxious differences for two different scales. With reference to Equation 5.8, x is the difference between noxious and non-noxious metrics for the first, unmodulated scale ($x = T^{scale=1}_{noxious} - T^{scale=1}_{nonnoxious}$). y is the difference between noxious and non-noxious metrics for the second, modulated scale ($y = T^{scale=2}_{noxious} - T^{scale=2}_{nonnoxious}$). I calculate statistical significance of the difference using the p value for this paired t

test. The paired t-test p value represents the probability of observing the pairwise differences observed if the populations for x and y were the same.

In order to answer part (b) of the research question, I use DeLong's test (DeLong, DeLong, and Clarke-Pearson 1988) to compare AUROCs for noxious-to-non-noxious differentiation between scales. I calculate statistical significance of the difference using the p value for this test, using the calculation method proposed by Sun and Xu (Sun and Xu 2014).

5.3 Results

5.3.1 Response magnitude estimate

A paired t-test shows that the response magnitude estimate is significantly higher, on average, after noxious stimuli than after non-noxious controls. This is the case when comparing noxious heel lances with tactile control heel lances ($t = 3.93$, $p < 0.001$, Figure 5.6a), and also when comparing noxious experimental stimui with background resting-state ($t = 7.59$, $p < 0.001$, Figure 5.7a). When used to differentiate noxious from non-noxious stimuli, it has an area under the Receiver Operating Characteristic curve (AUROC) of 0.673 (heel lances to tactile controls, Figure 5.6b) and 0.566 (noxious experimental to background, Figure 5.7b). This describes the ability of the template coefficient to differentiate between noxious and non-noxious stimuli, where an AUROC of 1 would perfectly predict the correct group for every stimulus, and an AUROC of 0.5 would be no better than chance.

5.3.2 Experimental stimuli

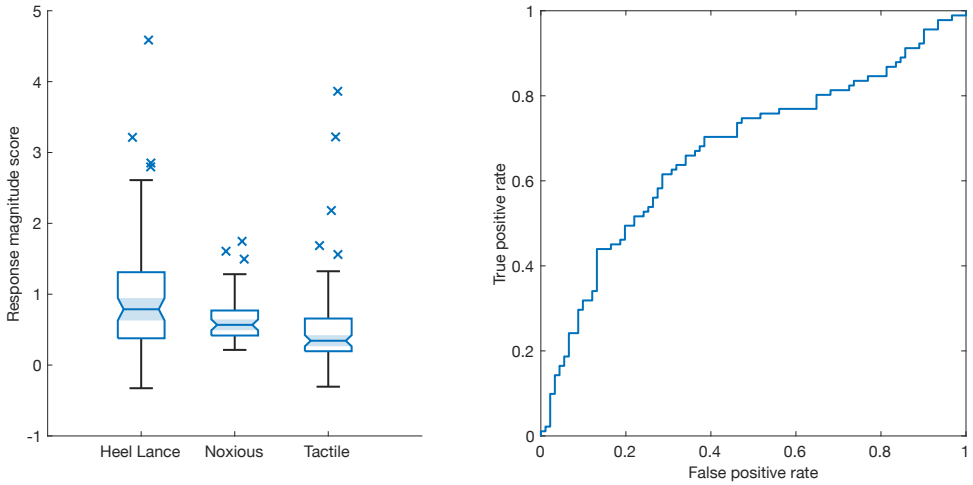


Figure 5.6: Comparison of response magnitude estimate for noxious and non-noxious stimuli, with no modulation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=3.93 (heel lance to tactile). Bars across each box show sample medians, with shaded notches indicating the 95% confidence interval for that median. Boxes whose notches do not overlap significantly differ in their sample medians ($p<0.05$). b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.67.

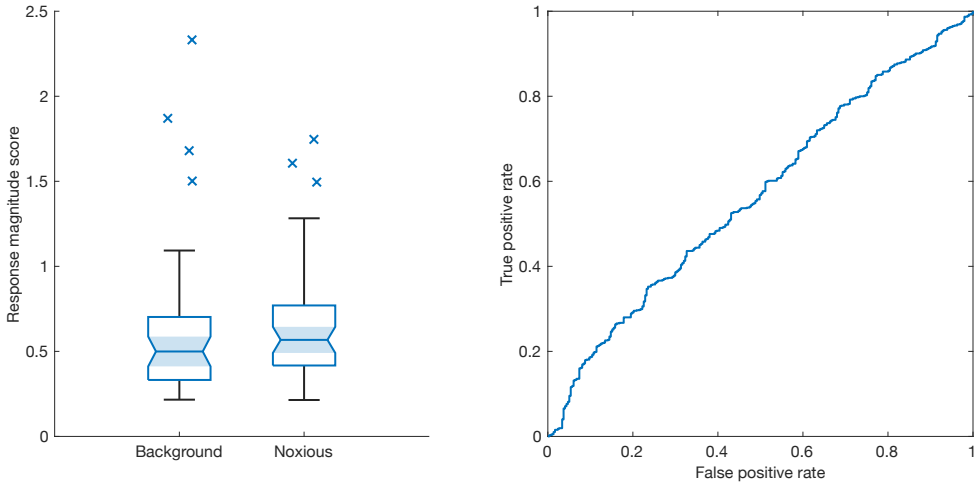


Figure 5.7: Comparison of response magnitude estimate for noxious stimuli and background, with no modulation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=7.59. Bars across each box show sample medians, with shaded notches indicating the 95% confidence interval for that median. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.57.

We now look at using multiple, experimental stimuli. Using experimental stimuli allows for multiple stimuli to be applied to a single infant in a single session. We can average response magnitude estimates from multiple stimuli in order to more closely approximate the long-run average. Absolute error is then calculated, compared to the average of all available responses for the infant as described in Equation 5.3. Error of the cumulative average of multiple stimuli is shown in figure 5.8. The error drops below 20% of the long-run average after 3 stimuli. With a pre-stimulus, post-stimulus, and inter-stimulus interval of 10 seconds, this will take approximately 40 seconds to apply the stimuli.

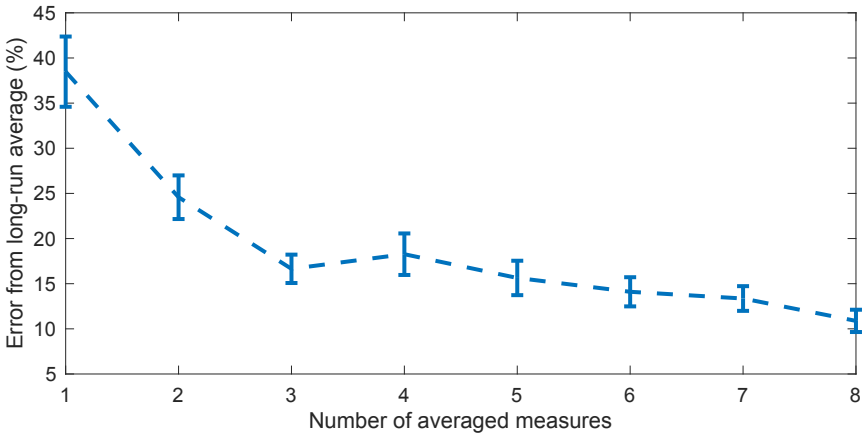


Figure 5.8: Train of averaged stimuli compared to long-run average. For each number of stimuli included in the average (x-axis), the average is compared to the long-run average (average of all available stimuli for that subject). The error shown here is the average of the absolute error values compared to long-run average.

5.3.3 Fit normalisation

Normalising the response magnitude to the Pearson's correlation of the response reduces the estimated magnitudes of trials with less-than-perfect fits. Normalising to fit in this way decreases the magnitude of the difference between noxious and non-noxious stimulation ($t = 3.48$, Figure 5.9a) and between noxious stimulation and background ($t = 7.17$, Figure 5.10a), although the difference is still significant ($p < 0.001$). AUROC remains approximately the same (0.65 and 0.57 respectively, Figures 5.9b and 5.10b).

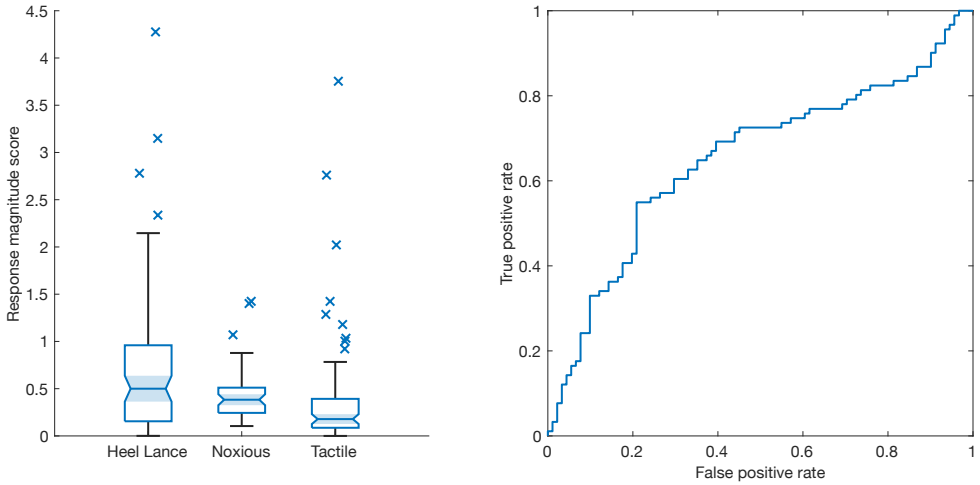


Figure 5.9: Comparison of response magnitude estimate for noxious and non-noxious stimuli, with fit normalisation. Bars across each box show sample medians, with shaded notches indicating the 95% confidence interval for that median. Boxes whose notches do not overlap significantly differ in their sample medians ($p<0.05$). a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=3.47 (heel lance to tactile). b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.652.

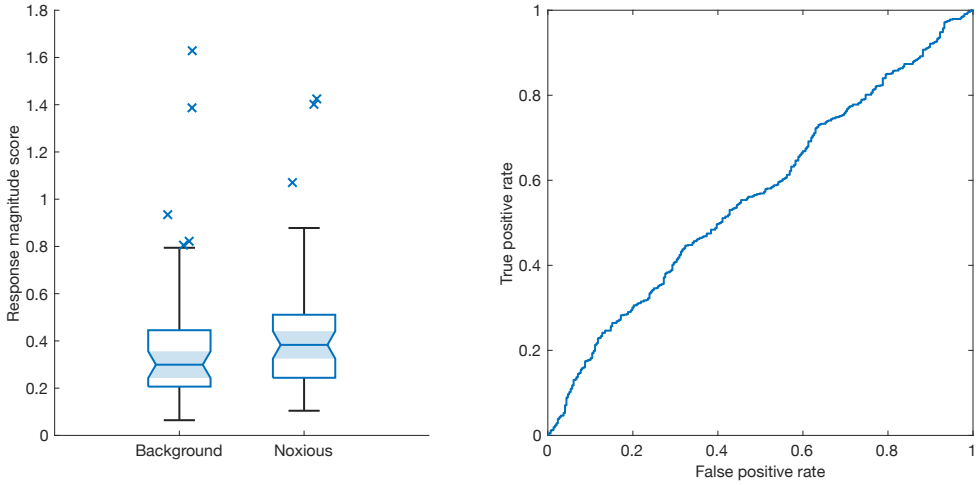


Figure 5.10: Comparison of response magnitude estimate for noxious stimuli and background, with fit normalisation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=7.17. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.566.

5.3.4 Baseline standardisation

Baseline standardisation estimates a distribution of baseline response magnitudes and uses this to z-score the estimated response magnitude to stimulus. This decreases the difference between noxious and non-noxious stimulation ($t = 2.91$, Figure 5.11a)

and between noxious stimulation and background ($t = 3.24$, Figure 5.12a). The difference is still significant ($p < 0.001$). AUROC remains similar (0.64 and 0.56 respectively, Figures 5.11b and 5.12b).

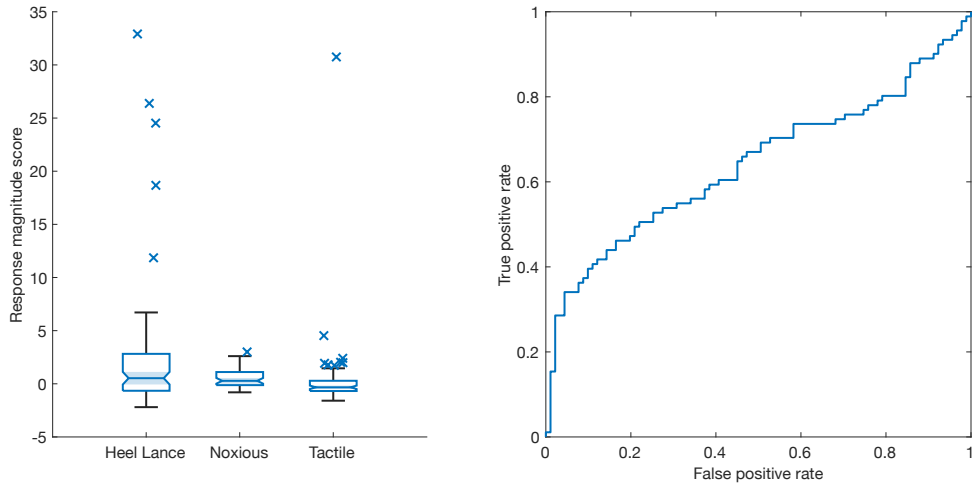


Figure 5.11: Comparison of response magnitude estimate for noxious and non-noxious stimuli, with baseline standardisation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=2.91 (heel lance to tactile). b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.64.

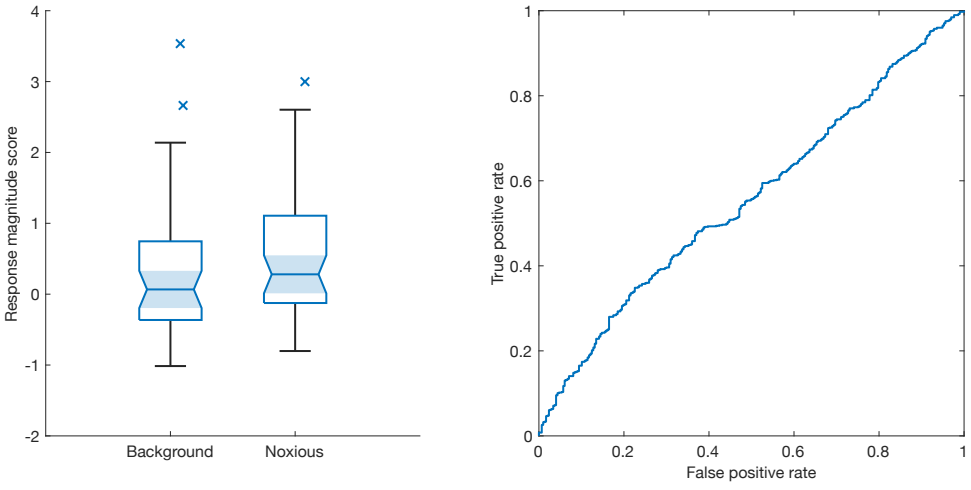


Figure 5.12: Comparison of response magnitude estimate for noxious and non-noxious stimuli, with baseline standardisation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=3.24. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.56.

5.3.5 Heartrate modulation

Heartrate modulation uses measurements from a separate modality - heartrate. The heartrate change (peak post-stimulus heartrate compared to average heartrate at baseline) is used to increase or decrease the estimated response magnitude metric. Experimental noxious stimuli were not assessed because stimuli in that dataset were recorded closer together, and the longer epochs required for heart rate analysis (±15 seconds) cause epochs to overlap.

The method for heartrate modulation is described in Equation 5.7 and illustrated in Figure 5.4. The difference between noxious and non-noxious stimulation is increased ($t = 5.21$, Figure 5.13a). The AUROC is also increased (0.71, Figure 5.13b).

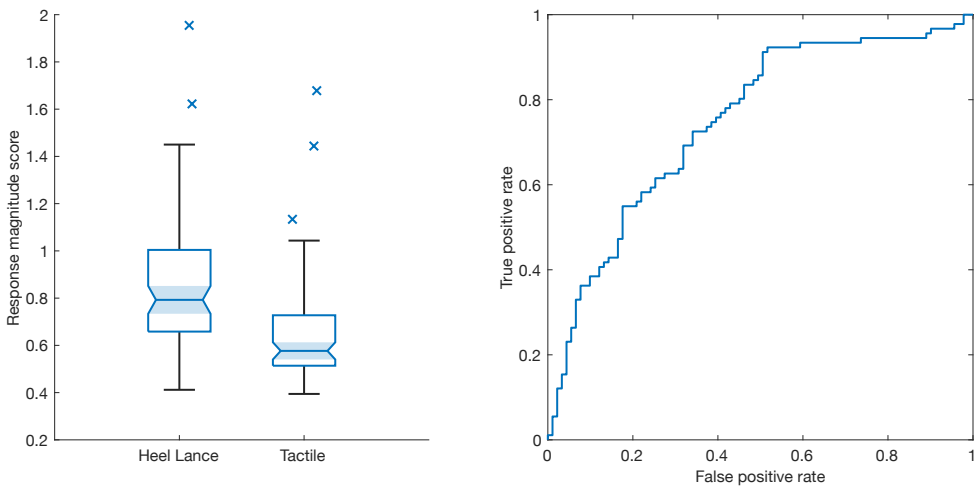


Figure 5.13: Comparison of response magnitude estimate for noxious and non-noxious stimuli, with proportional heartrate modulation. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances, noxious experimental stimuli and non-noxious tactile controls. t=4.68 (heel lance to tactile). b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.75.

5.3.6 Evaluation using an independent validation dataset

Previous results in this section have used datasets which I used for development of the different scales. Testing the known-group validity using an independent, validation dataset ensures that the results are generalisable. I used Dataset 2 (described in Chapter 3) as a validation dataset. For each scale I report two summary statistics, t-statistic and AUROC. In order to answer the research question I also

assess differences between scales using p value of the inter-scale paired t-test, as described in the Methods section.

Figure 5.14 shows the original scale with no modulation (t=3.35, AUROC=0.67). Figure 5.15 shows a scale using fit normalisation (t=2.99, AUROC=0.64). The test for difference from the unmodulated scale shows no significant difference (p=0.14 for t-statistic, 0.17 for AUROC). Figure 5.16 shows a scale using baseline standardisation (t=3.43, AUROC=0.68). The test for difference from the unmodulated scale shows no significant difference (p=0.29 for t-statistic, 0.68 for AUROC). Figure 5.16 shows a scale using heartrate modulation (t=5.11, AUROC=0.71). The test for difference from the unmodulated scale shows no significant difference (p=0.07 for t-statistic, 0.65 for AUROC). Table 5.2 summarises these results.

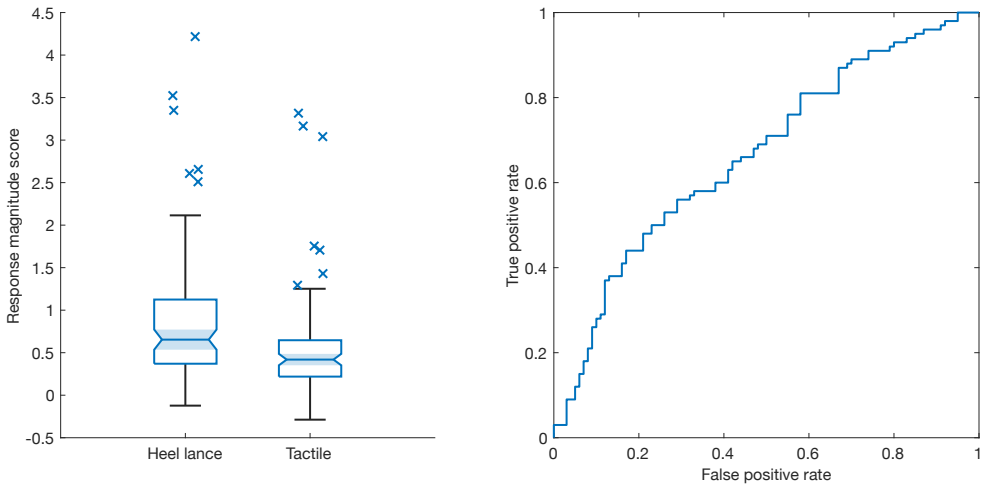


Figure 5.14: Comparison of response magnitude estimate for noxious and non-noxious stimuli, on the validation Dataset 2. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances and non-noxious tactile controls. t=3.35. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.67.

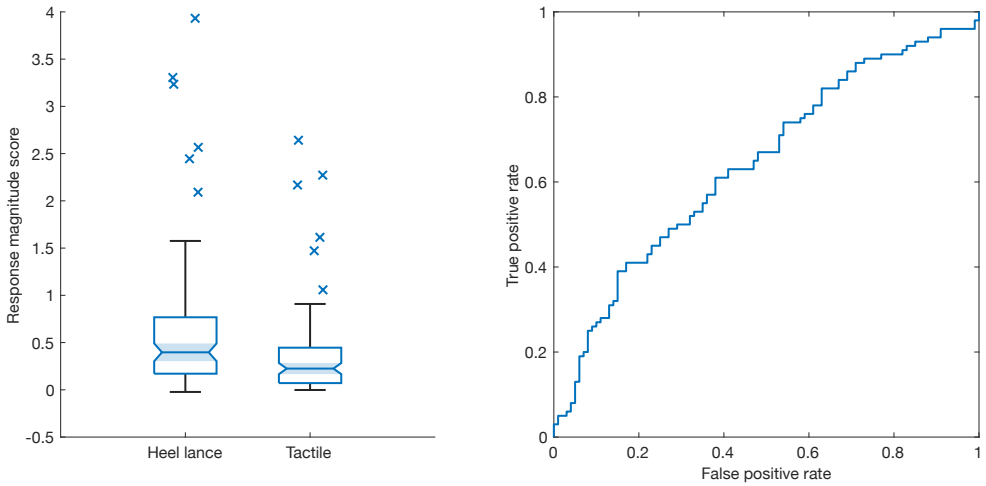


Figure 5.15: Comparison of response magnitude estimate for noxious and non-noxious stimuli, using fit normalisation on the validation Dataset 2. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances and non-noxious tactile controls. t=2.99. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.64.

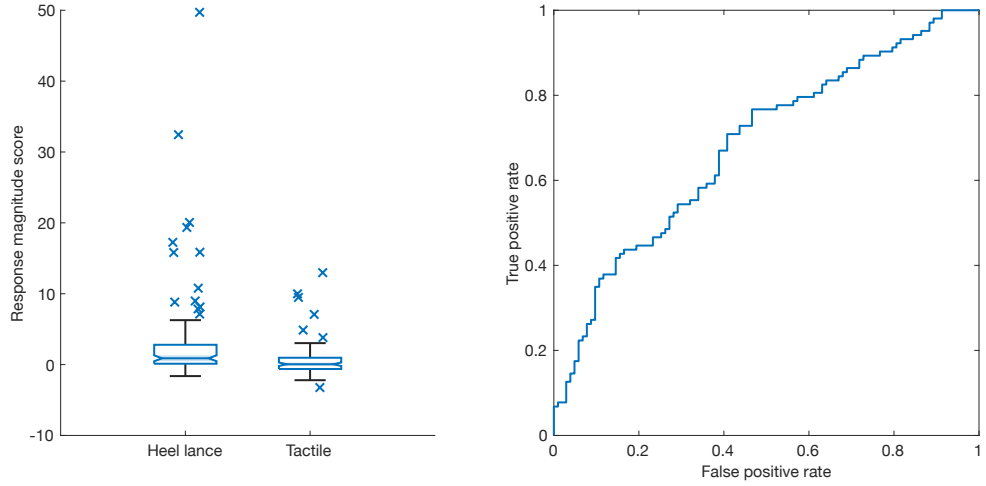


Figure 5.16: Comparison of response magnitude estimate for noxious and non-noxious stimuli, using baseline standardisation on the validation Dataset 2. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances and non-noxious tactile controls. t=3.43. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.68.

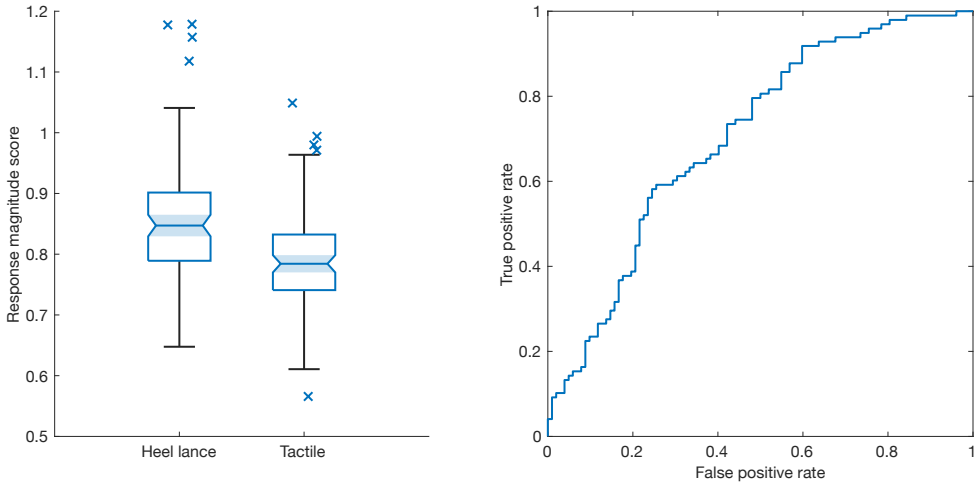


Figure 5.17: Comparison of response magnitude estimate for noxious and non-noxious stimuli, using heartrate modulation on the validation Dataset 2. a) Box-and-whisker plot comparing estimated response magnitudes for noxious heel lances and non-noxious tactile controls. t=5.11. b) ROC curve differentiating noxious heel lance from tactile control. AUROC=0.71.

Modulation	**t-statistic**	**p-value**
Unmodified	3.35	N/A
Fit normalisation	2.99	0.14
Baseline standardisation	3.43	0.29
Heartrate modulation	5.11	0.07
	AUROC	
Unmodified	0.67	N/A
Fit normalisation	0.64	0.17
Baseline standardisation	0.68	0.68
Heartrate modulation	0.71	0.65

Table 5.2: Performance summary of modulations of the response estimate. Only results for the validation dataset are shown here. p-value refers to the p value of the between-differences t test (for t-statistics) and the Delong's test (for AUROCs).

5.4 Discussion

The aim of this chapter has been to assess the known-groups validity of four proposed modifications to a scale of infant noxious response. I do this using noxious and non-noxious known groups. I evaluated the difference between each proposed scale with the unmodified scale using an independent dataset. This ensures that the results are generalisable, and the results which I discuss here all relate to the analysis on the independent dataset. I found no scales which improved known groups validity to a statistically significant extent.

In the clinical use context, the scale can be used as a diagnostic biomarker whereby the magnitude of nociceptive response to an event (such as a needlestick for inoculations or a heel lance for taking blood) is identified. Alternatively it can be used as a monitoring biomarker whereby regular measurement of noxious response (to stimuli applied for this purpose) allows monitoring of an infant's susceptibility to nociceptive response. In order to evaluate the measure for both of these types of biomarker, it was evaluated on two types of stimulus. For use as a diagnostic biomarker, measures on individual clinical heel lances were evaluated (compared to tactile control heel lances). For use as a monitoring biomarker, measures on a train of three experimental noxious stimuli were evaluated (compared to stimulus-free background).

The starting point to improve upon was the EEG 'template' method of Hartley et al. (2017). This method uses Singular Value Decomposition and Woody filtering to estimate a response magnitude to a known stimulus. The Singular Value Decomposition fits a template to the signal. The Woody filtering translates the template to each time-point in the range $300ms < t < 500ms$ and finds the point of highest cross-correlation between template and signal. Limiting the range of possible latencies is important in order to limit the result to biologically-plausible values, and in this chapter I treat the extent of this range (the level of 'jitter') as fixed. However, this jitter can have other values (and often does for very similar applications; cf. Hartley et al. 2017; Hauck et al. 2024), and the biologically-plausible range may vary with variables such as stimulus type, infant age, and stimulus location (Cobo et al.

2021). The jitter used here may not be optimal for this dataset - for example, see Figure 5.3c. Not only this, but this method requires an implicit prior assumption that all latencies within the range $300ms < t < 500ms$ are equally plausible, and latencies outside this range are impossible. This is a limitation of the method, and the full exploration of this is outside the scope of this chapter.

For each proposed modulation to the response estimate, t-statistic and AUROC. I calculated the p-value of the paired t-test between scales in order to answer the research question. I use a threshold of 0.05 to indicate statistical significance. This threshold is used because it is the standard. Traditionally a p value of less than 0.05 is considered 'statistically significant' (Kennedy-Shaffer 2019), but there is some discussion over whether this is a sensible threshold (Wayant, Scott, and Vassar 2018) or whether such a threshold is useful at all (McShane et al. 2019). By the 0.05 threshold, none of the modulations are statistically significant.

A summary of these results is shown in Table 5.2. None of the modulations improve known-groups discriminability of the known groups to a significant extent. I will briefly discuss some possible reasons for the lack of observed effect in these results.

Fit normalisation might fail to improve separation of noxious from non-noxious responses because the Woody filter jitter was not optimised (see discussion above). For example, the jitter could be too narrow and so Woody filtering did not always find the best response fit, or too wide and so Woody filtering selected fits to non-noxious response sections of signal.

Fit normalisation scales using the correlation, which has possible values within the unit interval ($[0, 1]$) and so will only reduce the value of the response magnitude estimate - estimates with good fits are not increased. This reduces the mean value of the response magnitude estimate. The units of the response magnitude are arbitrary, being simply a scaling factor for the original response template. An absolute value of a response magnitude has no biological meaning in itself, only relative to other values. Scaling values in this way has no impact on the biological relevance of the measure.

Baseline standardisation modifies the scale for the amplitude of pre-stimulus EEG activity, on the assumption that the magnitude of an individual's EEG activity is correlated with the magnitude of the ERP. Previous studies have identified a positive

correlation between EEG power and ERP magnitudes in adults and adolescents (Intriligator and Polich 1995; Lazzaro et al. 2001). These studies are, however, in a different context and on a different population to the data studied in this chapter.

I used a 'baseline' period immediately prior to the stimulus, as it eliminates the need for the user to record a separate period of EEG for the baseline. In all epochs used, the baseline period is free from any applied stimulus. These results may not show benefit from baseline standardisation because of confounding factors within the baseline period. The period of time immediately pre-stimulus was chosen in order to best standardise it across different stimuli, but in this period of time there is often disturbance from activity such as handling. In the future a helpful approach might be to identify the best time period to use as a baseline period - this may be best performed using prospectively-collected data where handling during data collection is minimised.

Heartrate modulation of the response estimate does not show an improvement in separation of groups or ability to classify epochs as noxious or tactile. There is evidence that both the magnitude of post-stimulus ERP and the maximum heartrate post-stimulus are related to an infant's noxious response (Hartley et al. 2017; van der Vaart et al. 2019). The principle of creating a scale using both EEG and heart rate is that incorporating two measures (i.e. two sources of information about a construct) improves the scale's ability to approximate that construct (Mokkink et al. 2010). This is the case as long as the two measures are not perfectly correlated, as this would indicate that they are measuring the same underlying factor. Measurements from these two modalities (EEG and heart rate) differ between individuals, and the relationship between the two modalities differs between individuals. These differences may be due to differences in individual biology, due to individual functional differences, or due to measurement error. The 'heartrate modulation' scale does not differentiate between these.

One potential limitation here is the reduced dataset on which this was tested. Noxious experimental stimuli were not tested, in order to prevent the larger ECG epochs from overlapping. Method 2 (proportional modulation) was optimised using heel lance data, which could result in overfitting when assessed using the same

data. This means that the results for this particular method may present an overly optimistic picture. However, both methods improve the t-statistic and AUROC (regardless of the optimisation chosen) and so this does not change the conclusion that this modulation improves the metric.

One potential limitation of the heartrate modulation method used is the method by which heart rate is obtained. As described in the Methods, I have used an open-source peak-finding algorithm, `findpeaks` (Sadek et al. 2017), to estimate heart rate from ECG. The benefits of using this method are that it performs adequately compared to other methods (Shimauchi et al. 2021) and the code is available for use. It is not necessarily the best method available. There are other methods available for heart rate estimation from ECG (Dogan and Dogan 2023), and there is recent work investigating method development (Chauhan, Agrawal, and Sabherwal 2023) and comparing the performance of different methods (Liu et al. 2018; Shimauchi et al. 2021; Suliman et al. 2019). For the purpose of this chapter I make the assumption that data obtained by this method is adequate.

This work shows the difficulty in translating measures used in research settings into clinical practice. There is a need for quantitative and objective measures of noxious response in clinical neonatal settings (Hartley and Slater 2014; Vanhatalo and Lauronen 2006) in order to identify and support infants who are 'at risk' of lasting neurodevelopmental changes. This need is yet to be addressed by the behavioural, physiological, and neurophysiological measures which are currently available (Cowen et al. 2015; Glenzel et al. 2023). This work may be informative for future studies, both by identifying promising methods (heartrate modulation, optimisation of filter windows) and by eliminating methods which are less promising (fit normalisation, baseline standardisation). By investigating a number of possible methods for improving the clinical usability of neuroimaging noxious response estimates, this work will hopefully be helpful in progressing towards a clinical measure of infant nociception.

6 Discussion

Contents

6.1 Aims of this thesis 107
6.2 Contribution to the state of the art 109
6.3 Limitations and future work 110
6.3.1 Pain and nociception 110
6.3.2 Acute and continuous pain 111
6.3.3 Other measurement modalities 113
6.3.4 Regulatory burden of proof 114
6.4 Implications and impact 115

6.1 Aims of this thesis

This thesis has aimed to develop methods for ERP analysis of infant noxious responses which translate into clinical and research use.

We cannot measure pain directly in nonverbal individuals, which makes it difficult to know when an infant is in pain and how much pain they are in. In the absence of a ground truth we rely on behavioural, physiological and neurophysiological responses to noxious stimulus which are indirectly related to pain. Neuroimaging methods allow us to measure aspects of the brain itself, from which responses to noxious stimuli derive. In particular, electroencephalographic measurement of event-related potentials allows us to directly measure signals from the cortex itself.

The development of these indirect metrics such as ERP analysis is important because it may allow us to evaluate the effect of analgesics and other pain management tools, and to identify which hospitalised infants would benefit from these interventions. Many newborns are subject to clinical procedures which may cause pain (Carbajal et al. 2008; Cruz, Fernandes, and Oliveira 2016) and there is evidence that early-life exposure to noxious stimuli has a negative impact on neurological development (Victoria and Murphy 2016; Vinall et al. 2014) and with health outcomes later in life (Williams and Lascelles 2020). For this reason it is of vital importance that these methods are developed for use both in the context of clinical trials, and in the context of clinical care.

ERP analysis has limitations and in this thesis I have addressed two of them by the development of new analysis methods.

The first method which I developed is a tool for automatic identification of artefact in EEG epochs. The aim was for this tool to identify artefact in short-duration EEG epochs with a single EEG channel. I did this by calculating a number of literature-derived metrics and training a machine learning model using data from 7 experienced raters. The method which I developed achieves the goal, and does so with the same accuracy as manual rating by experienced raters (balanced accuracy 0.81) and much faster (~500ms per epoch).

The second method which I developed is a clinically-usable, multi-modal scale of infant noxious response. This scale can serve as a method for identifying and quantifying noxious response in neonates. The goal was to create a single scale of noxious response, and assess the known-groups validity of the scale. The intended use would be as a diagnostic or monitoring biomarker in clinical settings. I evaluated four possible scales by comparing noxious heel lances with non-noxious control heel lances, using t-statistic and AUROC to quantify the ability to separate and distinguish one group from another. Using heartrate change to modulate the measure improved the ability to distinguish noxious stimuli from background. The other methods which I tested - fit normalisation and baseline standardisation - did not. These results could be informative for future work aiming at creating a single respnse scale.

6.2 Contribution to the state of the art

Chapter 4 research question: For a) a held-out test set of EEG epochs, and b) a separate validation set of EEG epochs, is the balanced classification accuracy greater than chance?

Chapter 4 develops an automated artefact detection algorithm. This algorithm estimates the probability that there is artefact in any given epoch of EEG, which allows us to distinguish between 'clean' epochs and those containing artefact. Algorithms for classification of artefact in EEG signals have already been developed, for example based on ICA (Debnath et al. 2020; Fló et al. 2022), ASR (Kumaravel et al. 2022) and channel similarity (Gabard-Durnam et al. 2018). These existing methods require multiple channels, long recordings, or multiple trials. The work which I present here is unique in that it can be applied to single-channel, single-trial event-related potential studies (Marchant et al. 2024).

Chapter 4 answers the research question which I introduced in Chapter 1. The algorithm has a balanced classification accuracy of 0.81 in the held-out test dataset (Dataset 1), and 0.78 in the clinical trial dataset (Dataset 2).

Chapter 5 research question: For a) an acute somatic clinical procedure, and b) an acute somatic experimental procedure, is the difference (noxious minus non-noxious) in group-level mean of modified response magnitude estimates significantly greater than the group-level mean of unmodified response magnitude estimates?

Chapter 5 develops a scale from the noxious-evoked response and the heartrate, representing an infant's response to noxious stimulus. This scale incorporates information about the measured magnitude of the neurological response, confidence in that measurement, and the physiological response. Previous work has developed multi-modal methods for classifying noxious from non-noxious responses in infants (van der Vaart et al. 2019; Zamzmi et al. 2016, 2017). The scale produced in this chapter is novel in that it is a single scale representing noxious response. In order to validate the scale I used known-group validity.

Chapter 5 answers the research question which I introduced in Chapter 1. When modulating the scale using the heart rate statistic, the difference in modified response

magnitude (group-level mean) is significantly greater than for the unmodified response magnitude. For all other modulations of the scale, it is not greater than the unmodified magnitude.

6.3 Limitations and future work

6.3.1 Pain and nociception

One limitation of this work is the philosophical assumptions upon which it is built. As described earlier - in Chapter 2.1 - the experience that we usually describe as pain is both a sensory and emotional experience, and so the nociceptive component does not fully describe the experience of pain. In fact, our inability to pin down the experience of pain is even more difficult than this might suggest: the transduction of nociception into pain is complex and highly non-linear. With a given nociceptive input, a person might experience any type and magnitude of the unlimited varieties of pain, or simply experience no pain at all.

In a scenario where we are measuring nociception because of its relation to pain, this non-linearity between nociception and pain is somewhat suboptimal. The methods which are used and developed in this work are metrics of the nervous system's pain response, but they cannot be said to directly *measure* pain itself. There are also different varieties of pain - sometimes classified into nociceptive, inflammative, or pathological pain (Woolf 2010). In measuring response to a nociceptive stimulus we focus on only one of these classes (nociceptive). We cannot claim to have any reasonably complete measure for the pain that an infant feels, so we cannot achieve any goal of reducing pain for infants (at least in a philosophically strict way).

There is, however, substantial evidence that exposure to nociceptive stimuli in early life has an impact on physiological (Victoria and Murphy 2016; Williams and Lascelles 2020), neurophysiological (Ranger et al. 2015; van den Bosch et al. 2015), and cognitive development (Giordano et al. 2023; Vinall et al. 2014). We do not know whether the infants in these studies felt the emotional aspect of pain, but we do know that (in aggregate, as a group) they experienced nociception and later experienced the effects described in these studies. Identifying and measuring nociceptive response

is therefore worthwhile because of these later effects. We can measure an infant's response to nociception, and so reduce nociception by conducting clinical trials and by optimising clinical plans. Regardless of whether an individual infant feels the unpleasant experience of pain, we know that reducing infants' exposure to this nociceptive response can improve their later development.

One potential avenue for future work is to identify whether the magnitude of the noxious response is related to changes in infant development (physiological, neurophysiological or cognitive). A longitudinal trial could achieve this by tracking the number of noxious procedures that infants were subject to, measuring infants' noxious response to these procedures (for example using the EEG-based method from Hartley et al. (2017) which is used in this work), and comparing these to their developmental trajectories (for example by tracking behavioural scores such as in Giordano et al. (2023) or white matter volume as in van den Bosch et al. (2015)). This could help to establish a causal link between early life nociception and later development.

6.3.2 Acute and continuous pain

The work in this thesis has focussed on neurological response signals which are very localised in both space and time. Indeed the longest post-stimulus period for which I have used data is 10 seconds (Chapter 5). Previous work has shown the most easily-discriminated part of the acute noxious response to be localised within several seconds or several hundred milliseconds of the stimulus (Hartley et al. 2018; van der Vaart et al. 2019). When discriminating between noxious and non-noxious responses, it is important to measure the parts which give the best signal (where 'signal' here means the response which can be attributed to nociception). Removing non-contributing time points from our analysis improves the signal-to-noise ratio which otherwise would be reduced by time-points with little correlation to noxious response ('noise'). Therefore, there is a good rationale for taking this temporally-focussed approach where we are discriminating between states (noxious and non-noxious, or indeed artefact and non-artefact).

The argument above does not necessarily mean that the entirety of the response is contained within this time period, or that the components of pain are only present

here. In focussing entirely on this short period of time and discarding all other time-points, I could potentially be missing noxious responses elsewhere. For example, recent work identified specific EEG microstates which occur in infants following a clinical heel lance (Bucsea et al. 2023; Rupawala et al. 2023). In preterm infants the latest-onset of identified microstates (of seven) had an average onset of 902 ms in preterm infants and 756 ms in term infants (Bucsea et al. 2023). Rupawala et al. (2022) found that microstates started later and lasted longer in preterm infants compared to term infants.

Many potentially-noxious stimuli in a clinical setting cannot easily be localised to such an acute response. Throughout this thesis I have designed methods for clinical stimuli which can be easily time-locked to a single time-point, such as heel lances or inoculations. On the other hand many noxious clinical procedures are ongoing, such as surgery (and recovery), intubation, extubation, or neonatal assessments for retinopathy of prematurity. These ongoing forms of noxious stimulus can evoke pain-related responses in infants. Some research has already investigated these types of noxious clinical procedures: for example identifying a difference compared to baseline in behavioural pain scales post-surgery (Alencar et al. 2012), during intubation (Caldwell and Watterberg 2015), and after assessment for retinopathy of prematurity (Belda et al. 2004).

Future work could extend the use of EEG neuroimaging biomarkers to measuring noxious response to continuous stimuli such as intubation. I believe that this could be achieved in two ways. The first potential method is to apply the same types of methods shown here, at several time points across a period of interest. For example, experimental stimuli could be applied at several points during a continuous clinical procedure such as a retinopathy of prematurity examination. Such a method could have the benefit of using these same techniques, but the main drawback is that this would involve two separate noxious stimuli: acute experimental and continuous clinical stimuli. We do not currently know how acute noxious response interacts with the response to ongoing stimuli. For example, perhaps the responses are purely summative and the magnitude of the noxious response is the sum of the response magnitudes for each individual stimulus. Perhaps, on the other hand, the descending

modulatory system inhibits noxious response when a continuous noxious stimulus is present. An experiment such as this could help to discover the mechanisms for noxious processing and response in the infant nervous system.

The second potential method for a future study on continuous noxious stimuli is to develop a new measure for neural noxious response which does not have to be time-locked to a known stimulus time. This could be achieved by comparing EEG (or other neuroimaging) from a continuous, noxious stimulus with that of a non-noxious control. By identifying features of the signal which can differentiate noxious from non-noxious response, we can identify candidate features. These would have to be capable of variation over time, in order to capture information about a continuous stimulus. Some potential features have already been used in Chapter 4. For example, wavelet analysis (Haar 1910; Orhan, Hekim, and Ozer 2011) decomposes signals into a series of waves, and can be used to give both spectral and temporal information about a signal. Fractal dimensional analysis (Higuchi 1988; Katz 1988) is a measure of the complexity of a signal, and could be used on a sliding window to show variation in signal complexity over time. This method would have the advantage of allowing 'true' continuous monitoring, with a temporal precision only limited by the measurement itself. However, there is no existing work within the infant pain literature on this type of method, and so it would be somewhat speculative.

6.3.3 Other measurement modalities

In this thesis I have focussed primarily on measurements which are derived from EEG and ECG. In Chapter 4 I design an automated artefact detection algorithm for single-channel EEG. In Chapter 5 I use EEG and ECG to create the noxious response scales. In Appendix A I create a software pipeline for EEG data processing. The thesis uses data from those two modalities only.

There are other modalities available for infant pain assessment. I introduced a number of these modalities previously, in Chapter 2. Behavioural response has been used by a number of researchers as a surrogate measure for pain. For example the duration (Franck 1986) or frequency content (Fuller 1991) of cry, the facial expression (Hatfield and Ely 2015), or body movement (Holsti et al. 2005). Behavioural

measurements can be made manually by subjective assessment (e.g. Holsti et al. 2005) or automatically by video-based algorithm (e.g. Heiderich, Leslie, and Guinsburg 2015; Zamzmi et al. 2016). Other neuroimaging modalities, aside from EEG, have been used in infant pain assessment. Examples include near-infrared spectroscopy (e.g. Bartocci et al. 2006) and functional MRI (e.g. Duff et al. 2020). Finally, non-imaging physiological measures such as skin conductance (e.g. Munsters et al. 2012) have previously been used to identify infant response to noxious stimuli. Any of these modalities represent potentially helpful data sources from which to assess pain and noxious response in infants.

Future work may find it useful to incorporate more than one modality into a single, multi-modal scale which is representative of infant noxious response. Some previous work has already created multi-modal scales, for example by combining physiological and behavioural modalities (Fitzgerald and Walker 2009; Roué et al. 2018; Salekin et al. 2021b). My work in Chapter 5 combining information from both EEG and heart rate is another possible method to create a multi-modal scale.

6.3.4 Regulatory burden of proof

One of the limitations of the work presented here relates to the overarching goal of making EEG neuroimaging useful as a nociception biomarker. None of the methods presented here have yet been shown to meet the regulatory requirements necessary for them to be used as biomarkers in clinical settings, or as proof-of-effect in analgesic clinical trials. Biomarkers such as this meet the definition (from the International Standards Organisation) of a medical device: "instrument, apparatus, implement, machine, appliance, implant, reagent for in vitro use, software, material or other similar or related article, intended by the manufacturer to be used, alone or in combination, for human beings, for [...] investigation, replacement, modification, or support of the anatomy or of a physiological process [...]" (ISO 13485 2016). In order to be used for any of these things, most legal jurisdictions require that new biomarkers are explicitly approved as medical devices by regulators. The three largest medical device markets (Maresova et al. 2015) - the European Union, Japan and the Unites States - all set requirements in legislation and all approve

medical devices through a central government agency. This is regulated by the European Medicines Agency (EMA) for the European Union, the Pharmaceuticals and Medical Devices Agency (PMDA) for Japan, and the Federal Food and Drug Agency (FDA) for the United States.

If these methods are only used in research settings - for example in clinical trials - then many jurisdictions exempt medical devices from full regulatory approval. For example in the European Union, medical devices which are not placed on the market and which are only used for research purposes are exempt from full device and manufacturer registration as long as they comply with a smaller subset of requirements, including complying with the General Safety and Performance Requirements listed in Annex I of the EU Medical Devices Regulations (*Regulation (EU) 2017/745 of the European Parliament* 2017). However, in order for a clinical trial to be used as proof-of-effect for treatments (such as analgesics) the trial end-point must be acceptable to the regulatory body, and regulatory bodies are not given to accepting evidence from unapproved biomarkers. Therefore, the burden lies on us to prove that these neuroimaging methods are safe and effective biomarkers, even if used purely in research settings.

There is substantial space for future work in gaining regulatory approval for these methods, for use in both clinical trials and in a clinical setting. Some work has already begun in this direction. A recent study protocol by Baxter et al. (2023) proposes using an EEG biomarker in an individual participant data meta-analysis to identify noxious response in infant data from several research labs worldwide. The proposed data processing method is the method which I developed in Appendix A, and the study will use the software pipeline which I developed there. It is hoped that this will pave the way for these methods to be accepted as an end-point for clinical trials studying analgesic efficacy in neonates.

6.4 Implications and impact

The work demonstrated here has several implications for research into infant pain.

Chapter 4 invents a method for automatically identifying short, single-channel epochs of EEG which contain artefact. This is the first method of its kind which can accurately classify EEG in this way for such small segments of signal. Segments of this size are not commonly found in clinical practice, however they are used in neonatal neuroimaging research, for example in ERP studies (Hartley et al. 2015; Molfese 2000; Rupawala et al. 2022; Schmidt Mellado et al. 2022). The implications of this work will therefore be seen within neonatal ERP research, where fast and reproducible artefact identification promises to improve the reproducibility of studies. One possible way of increasing the impact of this work are to improve the availability of the software, and to this end I intend to incorporate it into the software pipeline presented in Chapter A. Another method for increasing impact is to use the posterior probability, or even individual feature values, instead of using a binary classifier as currently. This may provide a more nuanced metric (or set of metrics) for researchers to use. The National Institute of Health & Care Research has provided an award to enable me to improve the impact of this work.

Chapter 5 creates a multi-modal, usable and reproducible metric for measurement of infant noxious response in clinical settings. While noxious-response metrics based on event-related potentials (Hartley et al. 2017) and multi-modal measures (Roué et al. 2018; Salekin et al. 2021b; van der Vaart et al. 2019) exist, this is the first to be developed specifically for clinical application (Cowen et al. 2015) with the principles of usability and single-subject reliability in mind. This is also the only part of the work presented here which aims specifically to develop monitoring biomarkers (as distinct from the response biomarkers typically used in research studies and clinical trials). Clinical use of quantitative biomarkers such as this in neonates is some way further away than the current focus of research use but there is no scientific, medical or technical reason that the same principles cannot be translated into clinical practice (Cowen et al. 2015). The results shown here do not provide a metric which can immediately impact clinical practice, but do provide helpful information on which to build further effort to translate these methods into clinically usable metrics. Were we to do so, this would provide quantitative information to support clinical pain management decisions in infants (Hartley and Slater 2014). Translation into

clinical practice such as this is vital for our ultimate goal of supporting the early-life wellbeing and development of every infant.

Appendices

A

A data pipeline for event-related potential studies

Contents

A.1 Introduction 119
A.1.1 Pre-processing of event-related potential data 119
A.1.2 The need for an alternative approach 120
A.1.3 Objective 123
A.2 Data 123
A.3 Pipeline description 123
A.3.1 General architecture 123
A.3.2 Lab-specific data transformation 124
A.3.3 EEG pre-processing 125
A.3.4 Response metrics estimation 126
A.3.5 Artefact detection and reporting 127
A.3.6 User interface 127
A.4 Filter responses 131
A.5 Pipeline output visualisation 135
A.6 Discussion 139

A.1 Introduction

A.1.1 Pre-processing of event-related potential data

As described previously, electroencephalography is a safe, non-invasive method of detecting and recording brain activity which is commonly used in research. There

are many methods available to record and process EEG data, and the most suitable method for any given application is dependent on the desired analysis. For example, research studies use various durations of EEG data - from continuous, multi-day EEG to short event-related potential (ERP) recordings - and various numbers of signal channels - from hundreds of channels in high-density electrode nets to single electrodes.

Raw EEG recorded directly from a scalp is generally not suitable for processing and analysis, and the data usually needs to be transformed in several ways to prepare it for analysis. These steps are together referred to as pre-processing.Pre-processing of EEG data is a significant part of any study relying on EEG data. The methodology for this data preparation is not standardised between laboratories. Much of the data preparation and processing requires subjective decisions from the analyst, and so the steps taken can and do differ depending on the analyst conducting the experiment (Debnath et al. 2020) which reduces the inter-rater reliability of the biomarker. In addition to this, because there are no standardised steps the exact method used for EEG analysis is often not reported alongside published results (Bigdely-Shamlo et al. 2015). This limits our ability to replicate and reproduce findings, to combine or compare results from different laboratories, and to generalise findings to other areas. This lack of reproducibility is increasingly becoming recognised as an issue in EEG research (Bigdely-Shamlo et al. 2015; Debnath et al. 2020; Rodrigues et al. 2021).

Software pipelines for automated pre-processing of EEG data are a topic of interest for many neuroscience research groups, and several have been developed to suit the needs of particular participant groups and analysis types. Some of these are listed in table A.1. Automating pre-processing in this way improves the reproducibility and repeatability of research studies using these automated methods, speeds up the pre-processing steps and makes it easier for non-technical users to carry out analyses.

A.1.2 The need for an alternative approach

One application of EEG is in neonatal studies that test response to noxious stimuli (Green et al. 2019; Hartley et al. 2018). This application requires - at minimum - only a single channel during a single ERP recording. The Paediatric Neuroimaging Group

Pipeline	Paper	Github Code	Group	Artefact removal
HAPPE	Gabard-Durnam et al. 2018	/lcnhappe/happe	Paeds	W-ICA & MARA
CTAP	Cowley, Korpela, and Torniainen 2017	/bwrc/ctap	Adult	ICA blink removal
PREP	Bigdely-Shamlo et al. 2015	eegstudy.org/prepcode	Adult	Inter-channel
Automagic	Pedroni, Bahreini, and Langer 2019	/methlabUZH/automagic/wiki	Adult	EOG-regression + MARA
APICE	Fló et al. 2022	/neurokidslab/eeg_preprocessing	Paeds	PCA, W-ICA & MARA
MADE	Debnath et al. 2020	/ChildDevLab/MADE-EEG-preprocessing-pipeline	Paeds	Modified ADJUST
NEAR	Kumaravel et al. 2022	/vpKumaravel/NEAR	Paeds	ASR
APP	da Cruz et al. 2018	Not available	Paeds	ICA (custom)
EEG-IP-L	Desjardins et al. 2021	/BUCANL/EEG-IP-L	Paeds	AMICA
EPOS	Rodrigues et al. 2021	osf.io/cw5qv	Adult	ADJUST & MARA

Table A.1: Published EEG pre-processing pipelines. Github code indicates the host location of source code, i.e. github.com/<insert-text-here> for entries starting with a slash, otherwise simply the URL. Group indicates the subject group for which the published work primarily addressed: adult or paediatrics.

in Oxford uses this method extensively. This approach has several idiosyncracies which make existing pre-processing pipelines unsuitable or difficult to use for EEG collected and processed in this way.

- Pre-processing pipelines are usually designed with a particular research group's needs in mind, and so relying on one creates possible future risks. For example, some pipelines are minimally documented (Fló et al. 2022; Gabard-Durnam et al. 2018) and others have stopped being supported since original publication (Desjardins et al. 2021). By creating our own software, we can control our own ongoing user support.

- The group's use of recording annotations for epoch labelling is unusual, and necessary due to the use of custom devices for automatic annotation. Existing pipelines do not allow annotations to be used for epoching.

- Neuroscan acquisition equipment (Compumedics Europe, Freiburg, Germany) records EEG to a proprietary, binary file format (Neuroscan Curry). Open-source software (Pontifex n.d.) converts this to the EEGlab standard data format, and I have contributed to this software. Using it requires some non-standard operations which are not part of existing pipelines.

- Existing pipelines assume that many channels are available and use pre-processing methods accordingly, for example by using channel interpolation (Debnath et al. 2020; Pedroni, Bahreini, and Langer 2019; Rodrigues et al. 2021) or spatially-based artefact identification (da Cruz et al. 2018).

- Some existing pipelines are designed only for continuous EEG and do not support fully epoched data (Cowley and Korpela 2018; Cowley, Korpela, and Torniainen 2017).

I developed an automated pipeline for EEG data which can process very low numbers of channels (1 at minimum) and very short recordings per subject (1 ERP at minimum). It is designed to support the type of ERP collection and analysis methods used in infant pain research.

A.1.3 Objective

The objective of including the details provided in this appendix is to demonstrate an EEG data pre-processing pipeline software which I developed as part of my doctorate. I describe the conceptual structure of the software and some of the background behind the decisions made. I show exemplar outputs which demonstrate typical values arising from particular types of data.

A.2 Data

I use two datasets to demonstrate the pipeline. Both of these datasets are described in detail in Chapter 3 (Datasets 1 & 2). Dataset 1 contains a range of stimuli types and infant ages. It was used to make design decisions about the pipeline. Dataset 2 contains data from a randomised clinical trial (Hauck et al. 2024). This dataset was collected after the pipeline was designed, and the trial used this pipeline for the published analysis. Results from the two datasets are both provided.

A.3 Pipeline description

A.3.1 General architecture

The pipeline uses a modular structure (Figure A.1), where each module groups together a number of data processing steps which would normally be performed together (either sequentially or in parallel). Each module can be run independently, and a set of module options allows the user to modify processing parameters or disable steps as required. This modular structure allows a user to perform any (or every) step of data processing that would typically be performed in an ERP study, and also perform additional steps for particular contexts.

The modules in the pipeline are as follows.

1. **Lab-specific data wrangling**. This module loads raw data files and parses them into EEGlab .set files, a file format which is compliant with the Brain Imaging Data Standard (BIDS EEG-extension), and which allows the use of a variety of tools via the EEGlab software. This module also includes any

laboratory-specific pre-processing which the data collection protocols might necessitate.

2. **EEG pre-processing**. This module contains several of the most common steps which are used in a variety of ERP study types. These are re-referencing, re-sampling, filtering, epoching and baseline amplitude correction. Through the user-defined parameters each step can be included or excluded from the module individually in each run, as needed.

3. **Response metrics estimation**. This module is specific to the study of noxious-evoked potentials using the methods of the Oxford Paediatric Neuroimaging Group. It compares a segment of an EEG epoch to a specified 'template' in order to estimate the best fit and provide magnitude and fit metrics.

A.3.2 Lab-specific data transformation

The first step of any EEG data processing is transforming the data into a standardised format. The process for this can differ widely between research labs due to differences in equipment, data collection protocols, and data storage methods. I chose the EEGlab .set file format as the format as it is widely used in neuroscience, allows a user access to the various tools available as plug-ins, and is compliant with the EEG extension of the Brain Imaging Data Standard (Pernet et al. 2019).

For the Oxford Paediatric Neuroimaging Group I developed a module for lab-specific data-wrangling with the following steps.

1. **Data type conversion**. The Oxford Paediatric Neuroimaging Group uses data collection equipment from Compumedics NeuroScan, which saves data in the "curry" proprietary file format and data structure. To convert to EEGlab format, the module uses the "load-curry" EEGlab plug-in.

2. **Removal of unwanted events**. Some unnecessary events may be recorded - for example due to accidental activation of a trigger. The module allows the

user to optionally identify the annotations of interest and discard events which do not have the associated annotation.

3. **Compensation for event recording delay**. Some recording equipment introduces a small, fixed delay in the synchronisation between data and event times. The module allows the user to optionally identify event types to be moved and automatically adjusts the time of these events by a fixed, user-defined time period.

A.3.3 EEG pre-processing

Data pre-processing consists of the following steps. The module allows all operations, or any subset of operations, to be completed sequentially in the same run. The (sub)set of operations to be completed is defined by the user at the time of processing.

1. **Re-referencing**. If the signal is to be re-referenced to an alternative channel, this can be performed. The user can choose the appropriate reference electrode for their purposes, depending on the question that they wish to address.

2. **Re-sampling**. If the signal is recorded at sampling rates very much higher than the frequencies of interest then re-sampling to a lower rate can reduce the time taken for subsequent processing operations to complete, without meaningfully reducing the useful information contained in the signal. An anti-aliasing filter is applied whenever this operation is carried out. By Nyquist's Theorem the absolute minimum sampling rate is twice the highest frequency of interest, in order that the anti-aliasing filter does not remove signals within the frequency range of interest. Higher sampling rates than this are usually used in most signal processing applications, and a common rule of thumb is to maintain a sampling rate of 5-6 times the highest frequency of interest.

3. **Filtering**. Removing frequencies in which EEG is not usually found helps to improve the signal-to-noise ratio. Low frequencies are often associated with movement artefact and voltage "drift". A high-pass filter is usually applied with a frequency cut-off of 0.1-1Hz (the default for this module is 1Hz pass-band

edge and 0.5Hz 6dB cut-off). High frequencies are often associated with muscle activity and electrical interference. A low-pass filter is usually applied with a frequency cut-off of 30-70Hz (the default for this module is 30Hz pass-band edge and 33.75Hz 6dB cut-off).

4. **Epoching**. This operation is completed after filters have been applied, so that filter edge effects are less likely to impact data within epochs. This is because edge effects are limited to the time periods at the start and end of the time period being filtered, and usually epochs will not be required at these time periods. This technique of epoching after filtering the entire continuous dataset does increase processing time (as filtering time is related to the number of samples to be filtered) but it was decided that this was an acceptable trade-off for the reduction of edge effects.

5. **Baseline amplitude correction**. Large absolute signal amplitudes within an epoch are not helpful when displaying an EEG epoch, and the absolute average magnitude has no inherent utility. In order to make epoch display easier, this operation centres an epoch's signal amplitude on zero by simple subtraction of the mean.

Figure A.2 is a simplified flow-chart showing the order of these operations. A single infant is used as an example to show how the processed EEG data changes through this process.

A.3.4 Response metrics estimation

Epoched EEG data can be useful for identifying and measuring event-related potentials. By identifying signal components which are related to stimulus response, a template can be defined which can be compared with EEG signal in individual epochs. Metrics for this template comparison may give an estimate of response to a stimulus. This operation specifically focusses on estimation of noxious response metrics.

Previous work (Hartley et al. 2017; Slater et al. 2010a) has characterised a nociceptive EEG activity waveform, or template, which is a measure of infant

nociceptive brain activity that is evoked by acute noxious stimulation. Using this template, metrics for the noxious response can be estimated in the following way.

1. Apply a Woody Adaptive Filter to jitter the epoch to the position with the maximum cross-correlation with the template.

2. Fit a linear regression model between the template and the Woody-filtered epoch. The regression coefficient indicates the magnitude of the response.

3. Calculate Pearson correlation between the template and the Woody-filtered epoch. The correlation coefficient indicates the fit of the response.

A.3.5 Artefact detection and reporting

My work presented in Chapter 4 has created an automated method of detecting artefact in individual, single-channel epochs of infant EEG. It has been tested in infants from 28-43 weeks gestational age. The method achieves a balanced accuracy of 0.81, which is as good as manual review of data, and accuracy is not significantly related to infant age or type of stimulus. It can be used within the pipeline.

A.3.6 User interface

I developed a graphical user interface for the pipeline, to make it more easily usable for researchers. The interface has two parts: a graphical user interface for point-and-click use on single subjects (Figure A.3) and an interface based in Matlab's command window for larger datasets and more complex scripting.

EEGlab is the standard software for neuroscience, and allows software developers to modify it using its plugin system. I integrated the pipeline with the EEGlab software in this way. When a user installs the plugin, it adds a menu item which allows the user to perform pre-processing and response estimation using the graphical user interface. Performing any of the operations in the pipeline modifies the `EEG` data structure which EEGlab uses to store information about a study subject; the operation is also recorded in the subject's `history` variable. Common user input errors are also checked for, and the error is automatically corrected or the user prompted as appropriate.

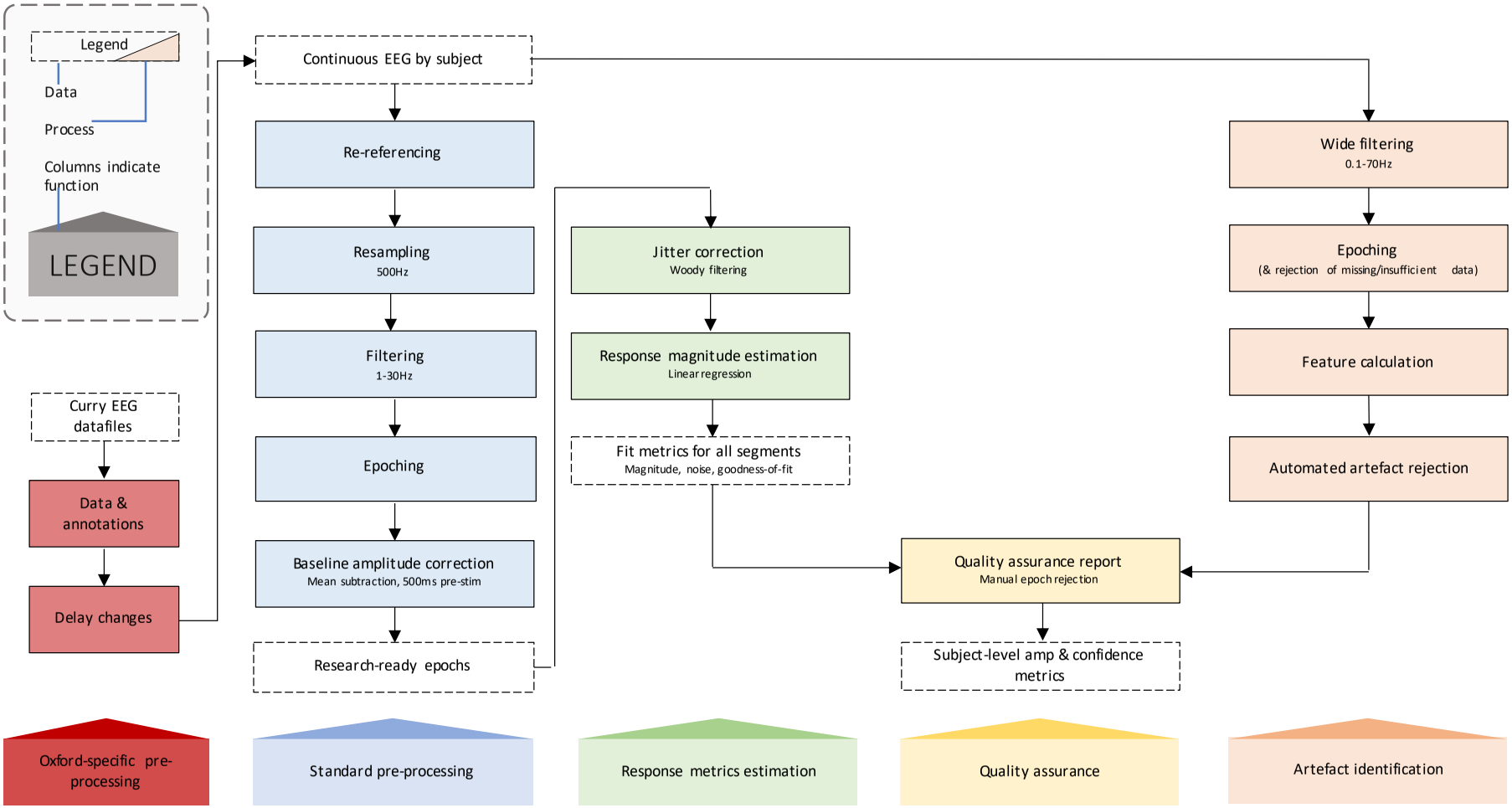


Figure A.1: Schematic of the EEG data pipeline. Arrows indicate the path of data through the processes, from curry EEG datafiles (far left) to subject-level amplitude and confidence metrics (bottom). Each box indicates a type of data (dotted outline) or a process (solid outline). In order to group processes which are commonly used together, processes are contained within separate functions. Each function has its own colour and column, as shown at the bottom of each column.

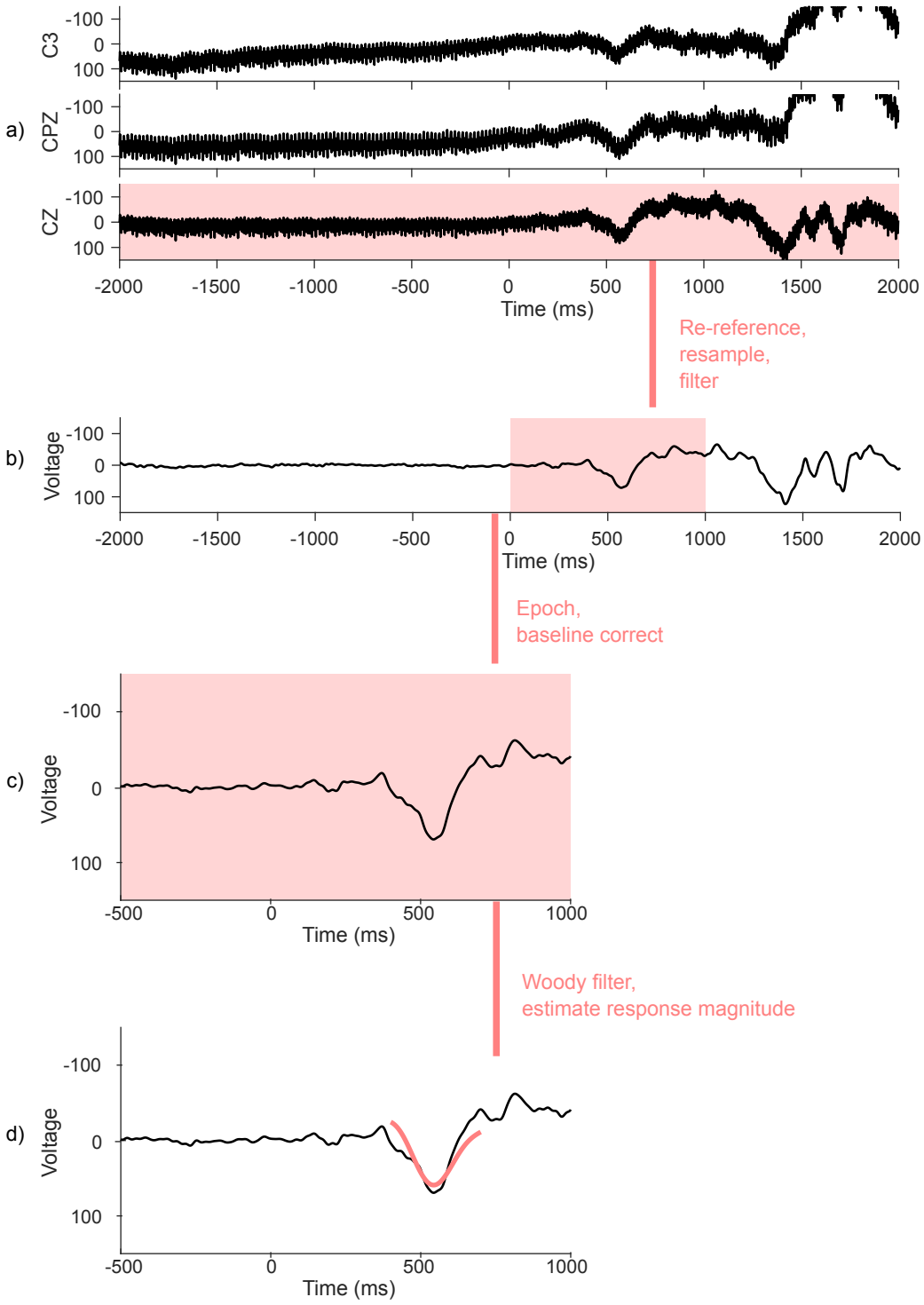


Figure A.2: A visualisation of the EEG for a single, term-age example infant. a) Raw, noisy data. 50Hz noise from mains interference can be seen. b) The shaded channel of *a* is re-referenced to Fz, resampled to 500Hz, and filtered using a 1-30Hz bandpass filter. c) The shaded epoch from *b* is extracted from the continuous signal, and baseline corrected. d) A response magnitude is estimated for the epoch.

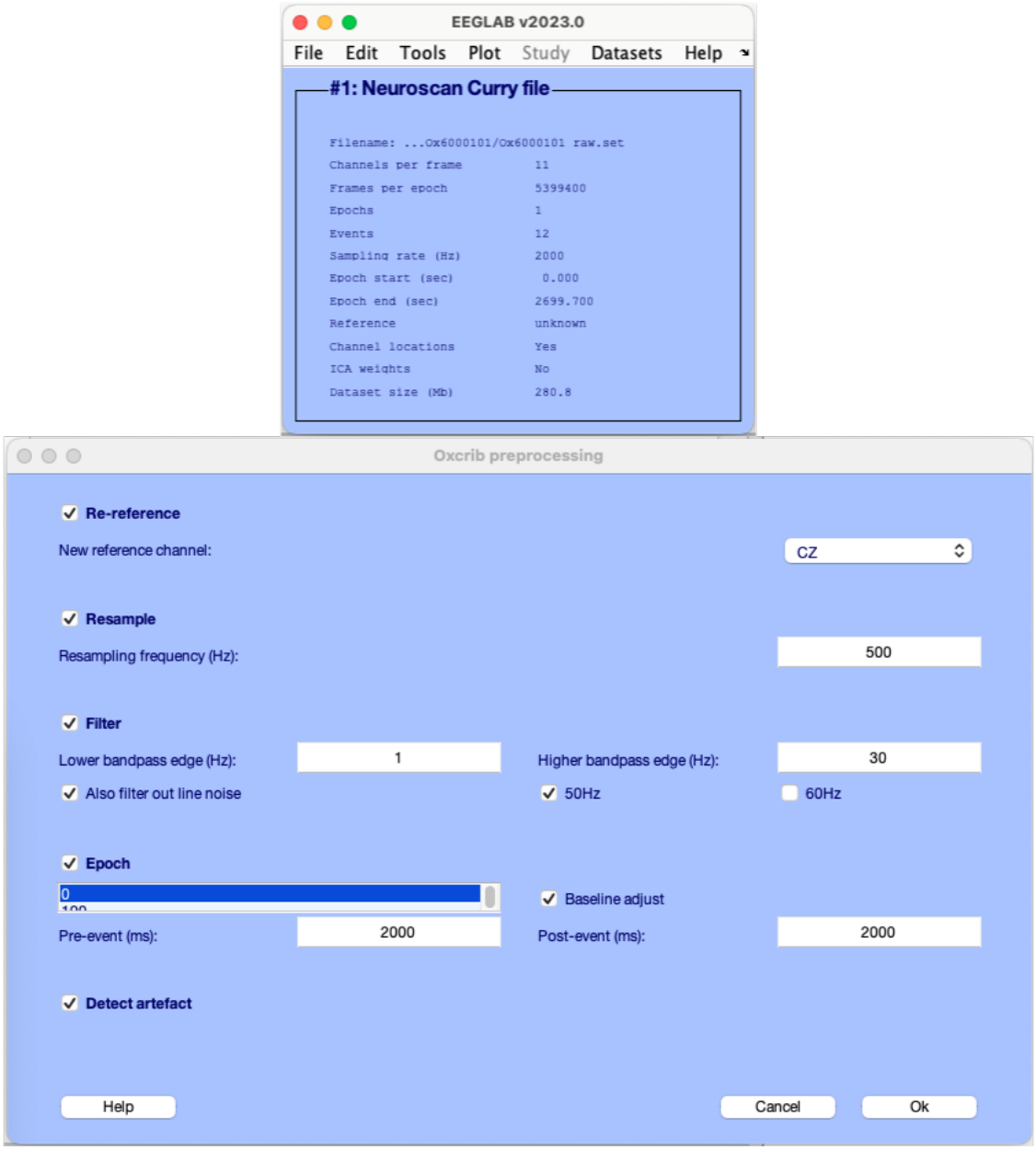


Figure A.3: A screenshot of the pipeline's graphical user interface, when accessed as a plugin within the EEGlab software.

A.4 Filter responses

Digital signal filters are a useful tool to remove gross artefact from recordings of physiological signals. In EEG signals, high-pass filtering is useful to attenuate slow drift such as skin conductance potentials (which can be aggravated by movement) and low-pass filtering is useful to attenuate high-frequency interference such as EMG or line noise. However, filters can cause distortions in the signal of interest. There are three main aspects of filter design to consider in ERP filtering. For each aspect, the decision on the approach to take is a trade-off, and so it is not possible to objectively identify the 'best' decision for every occasion.

Filter type is the first design decision. There are many types of filter, which can be broadly divided by their impulse response function and by their frequency domain characteristics (Smith 1997). An impulse response is a function's output, $f(x[n])$, given a unit-magnitude impulse input, $x = \delta[n]$. Filter designs with finite impulse response (FIR filters) are implemented by convolving the impulse response with the signal, and include moving average (strong time domain performance, weak frequency domain performance) and windowed-sinc filters (strong frequency domain performance). Filter design with infinite impulse responses (IIR filters) are typically implemented using a recursive system and include single pole (strong time domain performance) and Chebyshev filters (strong frequency domain response). EEGlab uses a Hamming windowed-sinc filter, which has a finite impulse response and a very fast transition band in the frequency domain. This allows it to separate frequencies extremely effectively, making it a good choice for filtering EEG.

This filter can be causal or non-causal. Filtering can be thought of as smoothing the signal across several time points. In causal filtering, the filter at any given time-point can only be influenced by signal at that point and earlier, not by later time-points (as that would break causality, hence be non-causal). This means that, for time-domain applications like evoked responses, the signal gets smeared out forward in time causing a latency shift. Non-causal filters attempt to undo this using both past and future time points to calculate the filter. One way of thinking of this is the filter being applied in both directions (forwards and backwards). The

benefit of non-causal filters, then, is that there should be no shift in the latency of a response. The downside is that non-causal filters are particularly susceptible to signal distortion at frequencies close to the filter cut-off frequency. This distortion is not limited by causality and so can cause distortions in a pre-stimulus baseline. Generally non-causal filters are used in ERP experiments, but this can cause pre-stimulus distortions if the filter cut-off is too close to the frequency (or frequencies) of interest.

The second design decision is frequency transition band: the range of frequencies within which the filter transitions from the passband (0dB, 1 gain) to the stopband (traditionally -3dB, 0.707 gain). A filter with a fast 'roll-off' has a narrow transition band, i.e. it transitions from passband to stopband within a short range of frequencies. The frequency response is defined by the filter's (time-domain) impulse response function, and so shorter transition bands (high precision in the frequency domain) have higher amplitude sinusoidal oscillations in the impulse response (low precision in the time domain). In addition, windowed-sinc filters - such as that used in EEGlab - modify transition band width by changing the length of the filter kernel. As the time required for signal convolution is proportional to the length of the signals, shorter transition bands (longer filter kernels) have longer compute times. For these reasons, there is a trade-off in the choice of transition band width (Smith 1997).

The third and final trade-off in filter design is frequency cut-off. This is typically defined as the frequency at which 50% of the signal power (3dB) is attenuated, although sometimes an alternative definition of 50% of signal amplitude (6dB) is used, especially in EEG (for historic reasons). A high-pass filter cut-off of 0.1 (6 dB) Hz is unlikely to introduce serious distortions to ERPs, whereas higher cut-offs are likely to introduce artefacts. The risk of artefact is higher with a higher high-pass filter, however low high-pass filters require a very narrow transition band and this has other consequences (as described above). Low-pass filters can have wider transition bands so the decision of cut-off frequency is less important. It is standard in EEG processing to use a 30Hz low-pass cut-off (Luck 2014).

To visually demonstrate the impact of filtering, filtered epochs are shown in Figure A.4. This uses Dataset 2, with a high-pass cut-off of 0.1Hz, 0.5Hz and no high-pass filter. As expected, filter design has little impact on grand averages

when there is no event-related response present. Grand average epochs containing event-related responses are impacted more by filter design. In particular, a positive voltage amplitude can be seen on the non-causal, 0.5Hz filter prior to the stimulus, possibly due to signal smoothing backwards in the time domain.

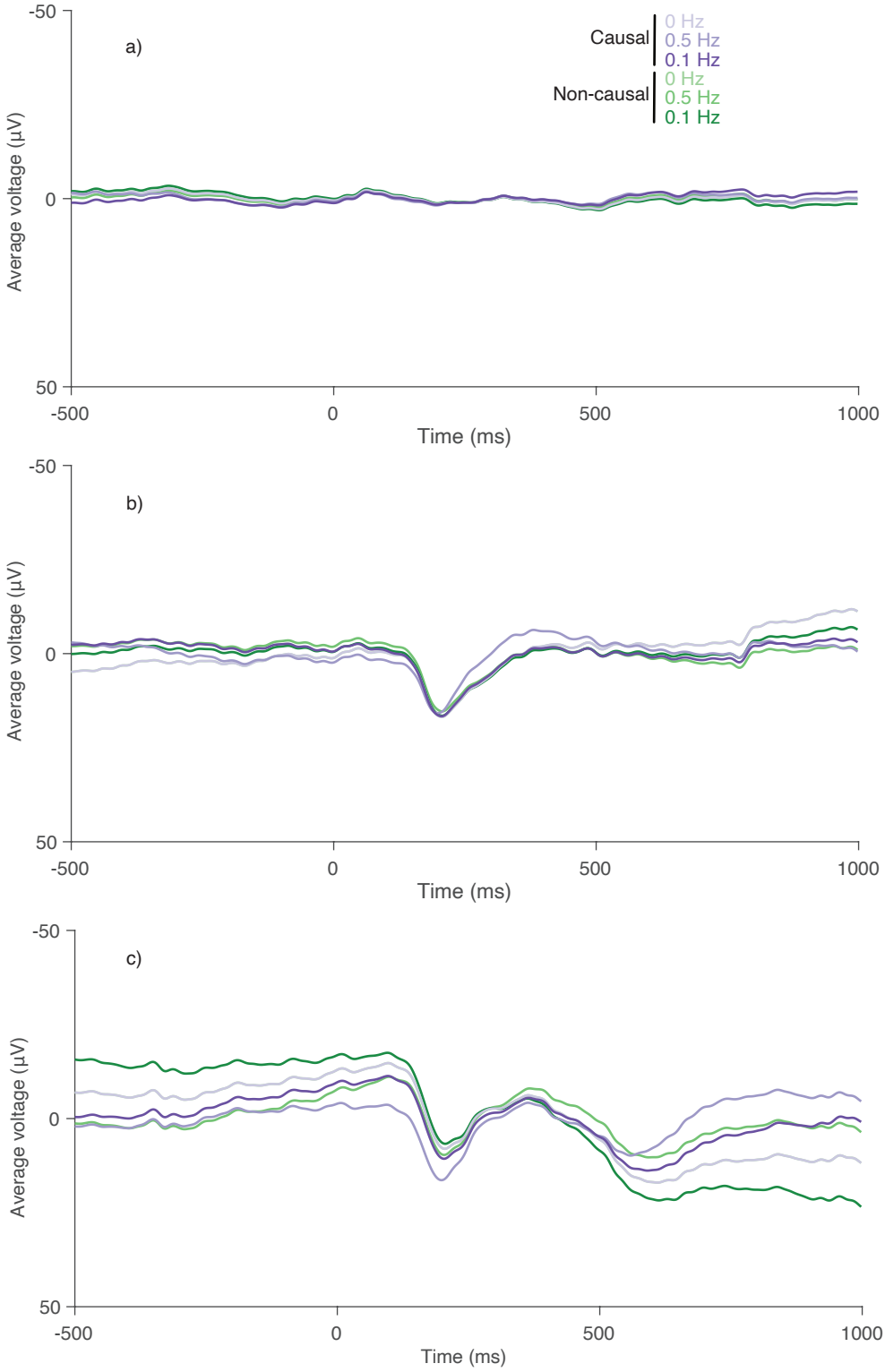


Figure A.4: Grand averages of EEG epochs from dataset 2, filtered using different filter designs and high-pass cut-off frequencies. Filtering improves signal-to-noise ratio by removing frequencies which contain no (or very little) signal of interest. This shows the impact of filter design on grand averages. Green lines indicate non-causal filter designs. Purple lines indicate causal filter designs. Colour saturation indicates high-pass cut-off: light is no high-pass, mid is 0.5Hz (6dB), and dark is 0.1Hz (6dB). a) Background epochs (n=106), b) Touch (control heel lance) epochs (n=106), c) Noxious (heel lance) epochs (n=88).

A.5 Pipeline output visualisation

For each epoch, the main numeric outputs of the pipeline are as follows.

- Noxious response magnitude estimate
- Correlation coefficient of response to template
- Posterior probability of clean epoch

For each dataset I visualise the distribution of these three values for all epochs in the dataset. For this I use Dataset 1 and Dataset 2 (see Chapter 3). A raincloud plot is used for visualisation, showing the probability density function in the upper part of each plot, and individual datapoints for each epoch in the lower part (Allen et al. 2021). To estimate probability density function a kernel smoothing function was used (matlab `ksdensity`) with a bandwidth of 0.1 (chosen empirically). Where data was bounded (i.e. for posterior probability) reflection boundary correction was used and the boundaries were expanded by 2% to remove boundary edge effects from the visualisation. Each dataset's epochs are split into three different stimulus types: background, noxious and vibrotactile. Epochs in dataset 2 can also be split by study site at which data was collected. Individual datapoints are also summarised by boxplots on the lower part of each plot showing mean and quartiles.

Figure A.5 shows output metrics from the pipeline when applied to dataset 1. Figure A.6 shows output metrics when applied to dataset 2. These results are useful to show distributions of output metrics which might be expected from datasets of these types. A notable exception to this is the posterior probability of a clean epoch in dataset 1 (Figure A.5c). This dataset is identical to the dataset used to train the artefact detector (described previously) and so posterior probabilities are clustered around $\mathbb{P} = 0$ and $\mathbb{P} = 1$ due to over-fitting on this data. This is expected but not generalisable to other datasets. It is included for completeness.

Figure A.7 uses the same data and methods as Figure A.6c but groups the epochs by trial site at which the data was collected.

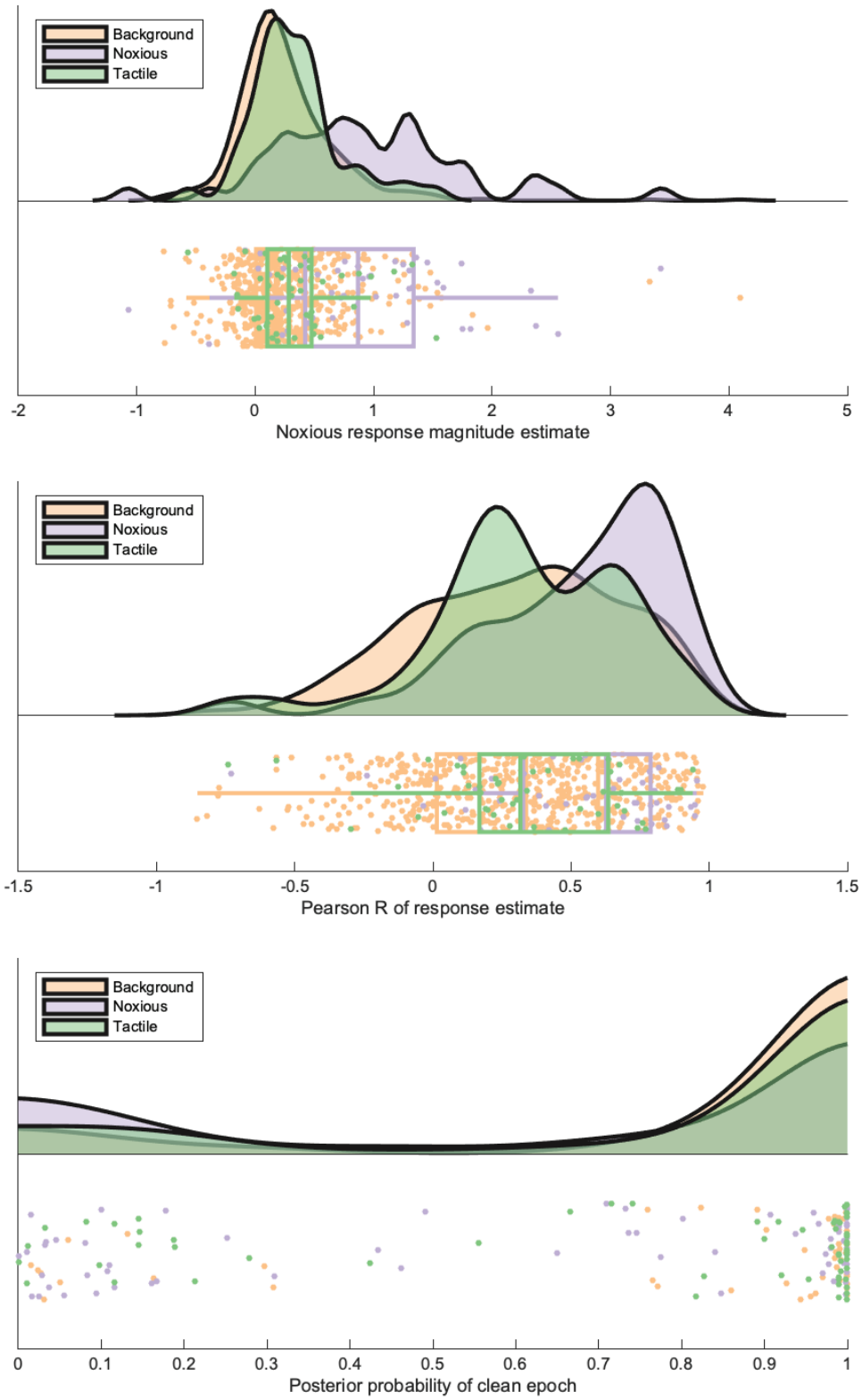


Figure A.5: Raincloud plots of three pipeline outputs, when the processing pipeline is applied to dataset 1. These show values that might be considered typical. Stimuli are background, i.e. no stimulus (n=541); noxious, i.e. clinical heel lance (n=68); and tactile, i.e. control heel lance (n=67). A) Response magnitude estimates. B) Correlation coefficient (Pearson's R) of the response magnitude. C) Estimated probability that an epoch is clean from artefact (The artefact model was trained on this dataset so over-fitting is visible).

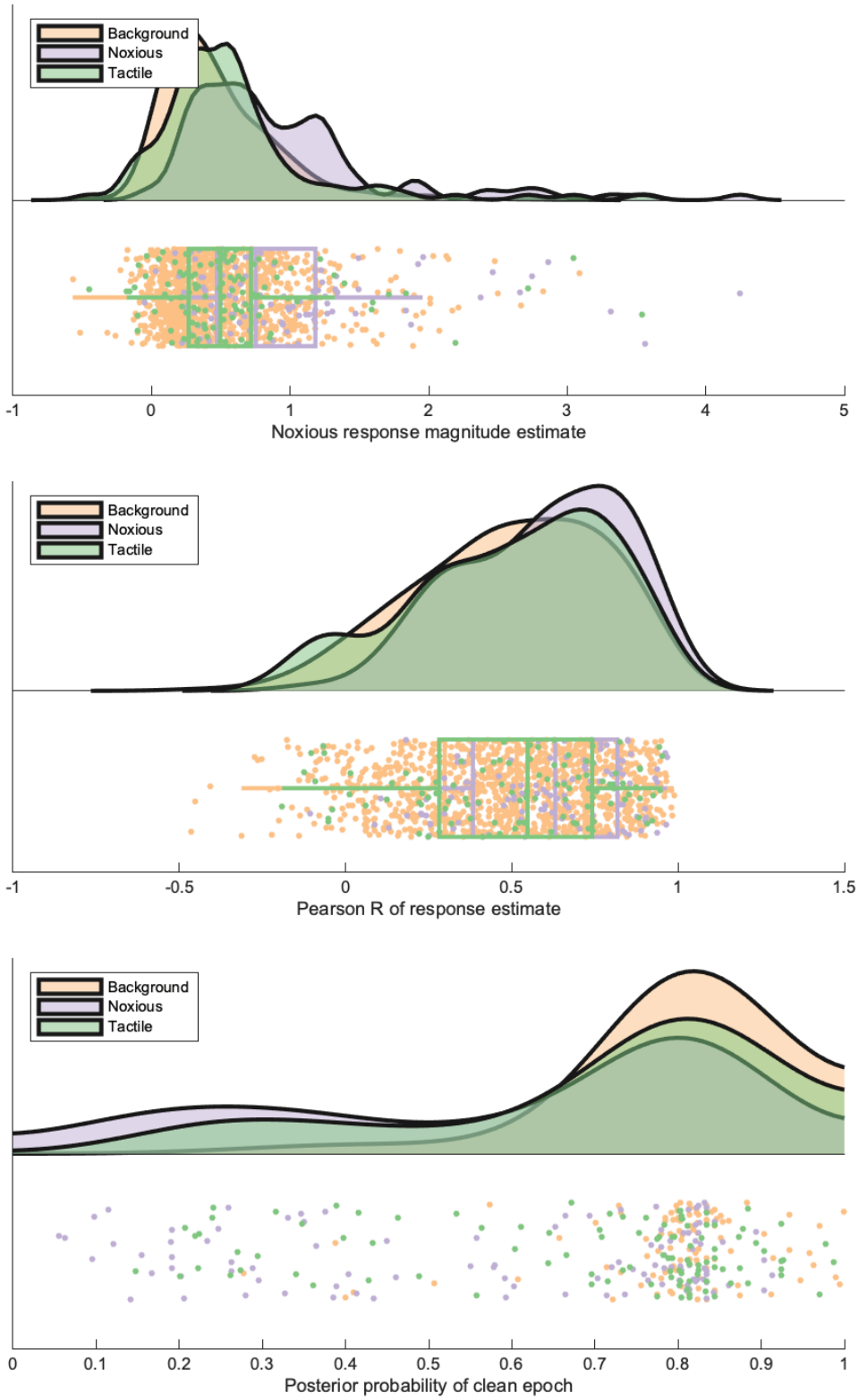


Figure A.6: Raincloud plots of three pipeline outputs, when the processing pipeline is applied to dataset 2. Stimuli at the epochs were background (n=1119); noxious (n=107); and tactile (n=107). A) Response magnitude estimates. B) Correlation coefficient (Pearson's R) of the response magnitude estimates after different types of stimuli. C) Estimated probability that an epoch is clean from artefact (this latter analysis uses fewer background epochs, n=107).

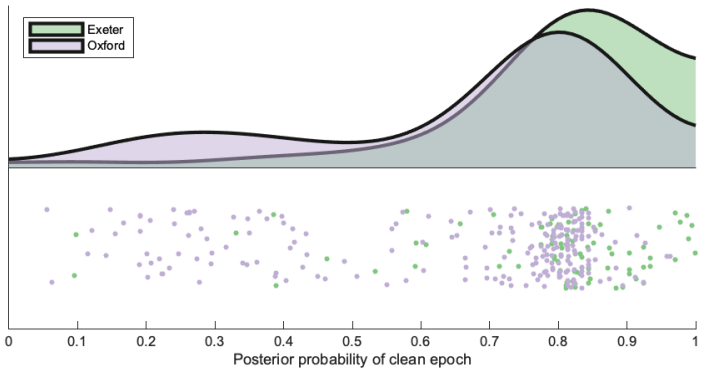


Figure A.7: Estimated posterior probability that a recorded epoch is clean from artefact when the processing pipeline is applied to dataset 2. All 3 types of stimulus are represented in the data. The epochs are split into those collected in the two study sites, Exeter (n=85) and Oxford (n=231).

A.6 Discussion

This work addresses the need for an automated data pipeline adapted to the specific needs of neonatal noxious-evoked potential studies. It describes the design and implementation this pipeline, and presents the products of this pipeline when applied to two separate sets of EEG epochs.

A wide range of tools are used for the processing of EEG data, with each research laboratory typically designing their own method which suits their specific experimental methods, data format, and analysis needs (Keil et al. 2014). The pipeline described here addresses the particular needs of the research group, and the aim has been to design a method which is as simple as possible to use while being generalisable within the range of studies usually conducted by the group. Alternative pipelines exist, many of which are highly cited and widely used in the field of paediatric EEG (Debnath et al. 2020; Gabard-Durnam et al. 2018). However they do not address this need for a lab-specific method.

The results presented here demonstrate exemplar values which might be calculated from noxious evoked potentials. Figure A.5 shows that when the pipeline estimates noxious response magnitudes after data pre-processing, there appear to be higher response magnitudes (A.5a) and high correlation coefficients (A.5b) for noxious stimuli than for tactile stimuli or background (no-stimulus) epochs. These two effects are expected, as previous studies have shown that this response appears to be related to noxious stimulation and not to tactile stimulation (Hartley et al. 2015, 2017). Statistical significance of this effect has not been calculated, however, as it is outside the scope of this work. This figure gives a visual approximation of the pipeline outputs which might be expected (assuming similar methods and participant characteristics) from sets of epochs containing noxious, tactile or no stimuli.

Figure A.7 shows artefact identified in each of the two sites of the Petal multi-centre RCT. Both have mean posterior probabilities around 0.8. There appears to be slightly more incidence of artefact (< 0.5 posterior probability that an epoch is clean) in epochs collected at the Oxford site, and slightly more incidence of high-certainty clean epochs (> 0.8 posterior probability that an epoch is clean) in epochs collected

at the Exeter site. Anecdotally, the clinical unit at the Exeter site is less busy and has less environmental noise around the infants. This may support the apparent difference, but numeric analysis is out of the scope of this work.

There are some limitations of the pipeline presented here. First, processing data from multiple study participants simltaneously is not currently possible using the graphical user interface, and so the user needs to use Matlab's command interface to achieve this (by using a *for* loop). As well as this, the artefact detection algorithm is currently implemented in Python and so is not fully integrated into the pipeline itself. Lastly, not all functionalities of the pipeline are documented. I intend to address these limitations in the future, and I have funding from the National Institute of Health Research (Oxford and Thames Valley Advanced Research Centre) to do so.

Aside from the limitations above, there is functionality that is not included in the pipeline described here. I have made this pipeline as a plugin for the popular EEGLAB software. This has commonly-used functionality to facilitate workflows such as batch processing, and so any user of the pipeline will have access to the full suite of standard tools available in that software.

In summary, this appendix demonstrates the data pre-processing software which I made. I used it in Chapter 5 of this thesis and it has since been used for a randomised clinical trial (Hauck et al. 2024).

B
Publications

Publications arising from the work in this thesis

A machine learning artefact detection method for single-channel infant event-related potential studies. Simon Marchant, Marianne van der Vaart, Kirubin Pillay, Luke Baxter, Aomesh Bhatt, Sean Fitzgibbon, Caroline Hartley, Rebeccah Slater. In: *Journal of Neural Engineering (2024).*

Patent: **Systems and methods for measuring a response of a subject to an event.** Rebeccah Slater, Aomesh Bhatt, Kirubin Pillay, Caroline Hartley, Simon Marchant. *Published as: WO2023084208A1.*

Related publications

The relationship between neonatal respiration and EEG-recorded brain activity: a systematic review. Fatima Usman, Simon Marchant, Luke Baxter, Hamisu M. Salihu, Muktar H. Aliyu, Eleri Adams, Caroline Hartley. In: *Clinical Neurophysiology Practice (2023).*

Statistical analysis plan for the Petal trial: the effects of parental touch

on relieving acute procedural pain in neonates. Luke Baxter, Annalisa Hauck, Aomesh Bhatt, Maria M. Cobo, Caroline Hartley, Simon Marchant, Ravi Poorun, Marianne van der Vaart, and Rebeccah Slater. In: *Wellcome Open Research (2023).*

Effect of parental touch on relieving acute procedural pain in neonates and parental anxiety (Petal): a multicentre randomised controlled trial. Annalisa GV Hauck, Marianne van der Vaart, Eleri Adams, Luke Baxter, Aomesh Bhatt, Daniel Crankshaw, Amraj Dhami, Ria Evans Fry, Marina BO Freire, Caroline Hartley, Roshni Mansfield, Simon Marchant, Vaneesha Monk, Fiona Moultrie, Mariska Peck, Shellie Robinson, Jean Yong, Ravi Poorun, Maria M Cobo, Rebeccah Slater. In: *The Lancet Child & Adolescent Health (2024).*

Establishing a standardised approach for the measurement of neonatal noxious-evoked brain activity in response to an acute somatic nociceptive heel lance stimulus. Marianne Aspbury, Roshni C Mansfield, Luke Baxter, Aomesh Bhatt, Maria M. Cobo, Sean P Fitzgibbon, Caroline Hartley, Annalisa Hauck, Simon Marchant, Vaneesha Monk, Kirubin Pillay, Ravi Poorun, Marianne van der Vaart, Rebeccah Slater. In: *Cortex (2024).*

C
Funding Acknowledgements

I would like to thank the following organisations for their financial support, without which this work would not have been possible.

- The Engineering and Physical Sciences Research Council
- The Alan Turing Institute
- The Reckitt Benckiser Group PLC
- The National Institute of Health Research

D

STROBE Checklists

This appendix contains checklists for Chapters 4 and 5 in the format required by the STROBE Statement for reporting guidelines on observational studies in epidemiology (Elm et al. 2007). The STROBE statement 'is a checklist of items that should be addressed in articles reporting on the three main study designs of analytical epidemiology: cohort, case-control, and cross-sectional studies' (Elm et al. 2007). I have chosen to use the checklist for cross-sectional studies, as it most closely matches the format of the work in this thesis. The thesis does not describe a cohort, case-control or cross-sectional study by the definitions used in STROBE (Vandenbroucke et al. 2007) and so I have mapped each item as closely as possible to the checklist's intention (Elm et al. 2007; Vandenbroucke et al. 2007). The appendix is included following an examiner request.

Chapter 4 STROBE	Item	Page Number
Title	1	47
Introduction		
Background	2	48-50
Objectives	3	50
Methods		
Study design	4	50
Setting	5	50-51
Participants	6	40-43
Variables	7	58
Data sources/measurement	8	58
Bias	9	51
Study size	10	50-51
Quantitative variables	11	51-59
Statistical methods	12	58-59
Results		
Participants	13	40-43
Descriptive data	14	59-62
Outcome data	15	62-67
Main results	16	62-67
Other analyses	17	61-62
Discussion		
Key results	18	68
Limitations	19	70-72
Interpretation	20	69
Generalisability	21	69
Other information		
Funding	22	143

Table D.1: STROBE checklist for Chapter 4. The STROBE Statement (Elm et al. 2007) recommends including a checklist of items that should be included in reports. Page numbers refer to the page of the thesis which includes each item.

Chapter 5 STROBE	Item	Page Number
Title	1	73
Introduction		
Background	2	75-81
Objectives	3	74
Methods		
Study design	4	83-84
Setting	5	83-84
Participants	6	42-44
Variables	7	85-92
Data sources/measurement	8	85-92
Bias	9	81
Study size	10	83-84
Quantitative variables	11	85-89
Statistical methods	12	92-94
Results		
Participants	13	42-44
Descriptive data	14	94-99
Outcome data	15	99-100
Main results	16	99-100
Other analyses	17	N/A
Discussion		
Key results	18	103-104
Limitations	19	104-106
Interpretation	20	106
Generalisability	21	103
Other information		
Funding	22	143

Table D.2: STROBE checklist for Chapter 4. The STROBE Statement (Elm et al. 2007) recommends including a checklist of items that should be included in reports. Page numbers refer to the page of the thesis which includes each item. 'Other analyses' is not applicable as only the main analyses were performed.

Bibliography

Abdelsami Awad, H et al. (Dec. 2018). "Analysis of Pain Effect on EEG Recordings and Oral Sucrose Suckling Effect on Pain Reduction in Neonates". In: *QJM: An International Journal of Medicine* 111.suppl_1. ISSN: 1460-2725, 1460-2393. DOI: 10.1093/qjmed/hcy200.158 (cited page 38).

Accardo, A. et al. (Nov. 1997). "Use of the Fractal Dimension for the Analysis of Electroencephalographic Time Series". In: *Biol Cybern* 77.5, pp. 339–350. ISSN: 1432-0770. DOI: 10.1007/s004220050394 (cited page 54).

Alencar, Ana Julia Couto et al. (Jan. 2012). "Efficacy of Tramadol versus Fentanyl for Postoperative Analgesia in Neonates". In: *Arch Dis Child Fetal Neonatal Ed* 97.1, F24–F29. ISSN: 1359-2998, 1468-2052. DOI: 10.1136/adc.2010.203851 (cited page 112).

Allegaert, Karel (Nov. 2017). "Better Medicines for Neonates: Improving Medicine Development, Testing, and Prescribing". In: *Early Human Development*. Special Issue: Neonatal Update 2017: "The Science of Newborn Care" 114, pp. 22–25. ISSN: 0378-3782. DOI: 10.1016/j.earlhumdev.2017.09.007 (cited page 19).

Allen, M et al. (2021). "Raincloud Plots: A Multi-Platform Tool for Robust Data Visualization [Version 2; Peer Review: 2 Approved]". In: *Wellcome Open Research* 4.63. DOI: 10.12688/wellcomeopenres.15191.2 (cited page 135).

American Academy of Pediatrics and Canadian Paediatric Society (Nov. 2006). "Prevention and Management of Pain in the Neonate: An Update". In: *Pediatrics* 118.5, pp. 2231–2241. ISSN: 0031-4005, 1098-4275. DOI: 10.1542/peds.2006-2277 (cited page 6).

Amzica, Florin and Fernando H. Lopes da Silva (Nov. 2017). "C2Cellular Substrates of Brain Rhythms". In: *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Ed. by Donald L. Schomer et al. Oxford University Press, p. 0. ISBN: 978-0-19-022848-4. DOI: 10.1093/med/9780190228484.003.0002 (cited pages 19, 20).

Anand, K.J.S. and P.R. Hickey (Nov. 1987). "Pain and Its Effects in the Human Neonate and Fetus". In: *New England Journal of Medicine* 317.21, pp. 1321–1329. ISSN: 0028-4793. DOI: 10.1056/NEJM198711193172105 (cited page 15).

Apgar, Virginia (May 2015). "A Proposal for a New Method of Evaluation of the Newborn Infant". In: *Anesthesia & Analgesia* 120.5, pp. 1056–1059. ISSN: 0003-2999. DOI: 10.1213/ANE.0b013e31829bdc5c (cited pages 43–45).

Askari, Shahbaz et al. (Mar. 2021). "Lighting up Babies' Brains: Development of a Combined NIRS/EEG System for Infants". In: *Biophotonics in Exercise Science, Sports Medicine, Health Monitoring Technologies, and Wearables II*. Ed. by Babak Shadgan and Amir H. Gandjbakhche. Online Only, United States: SPIE,

p. 20. ISBN: 978-1-5106-4111-2 978-1-5106-4112-9. DOI: 10.1117/12.2595899 (cited page 38).

Aspbury, M. (2023). "Developing Neuroimaging Methods for Clinical Translation and Better Understanding Neonatal Brain Development". http://purl.org/dc/dcmitype/Text. University of Oxford (cited pages 33, 34).

Baarslag, Manuel A. et al. (Jan. 2017). "Paracetamol and Morphine for Infant and Neonatal Pain; Still a Long Way to Go?" In: *Expert Review of Clinical Pharmacology* 10.1, pp. 111–126. ISSN: 1751-2433, 1751-2441. DOI: 10.1080/17512433.2017.1254040 (cited page 18).

Balice-Bourgois, Colette et al. (May 2020). "A Systematic Review of Clinical Practice Guidelines for Acute Procedural Pain on Neonates". In: *The Clinical Journal of Pain* 36.5, p. 390. ISSN: 0749-8047. DOI: 10.1097/AJP.0000000000000808 (cited page 6).

Baliki, Marwan N. and A. Vania Apkarian (Aug. 2015). "Nociception, Pain, Negative Moods, and Behavior Selection". In: *Neuron* 87.3, pp. 474–491. ISSN: 08966273. DOI: 10.1016/j.neuron.2015.06.005 (cited page 12).

Barr, R. G (Sept. 1998). "Reflections on Measuring Pain in Infants: Dissociation in Responsive Systems and "Honest Signalling"". In: *Archives of Disease in Childhood - Fetal and Neonatal Edition* 79.2, F152–F156. ISSN: 1359-2998, 1468-2052. DOI: 10.1136/fn.79.2.F152 (cited page 75).

Bartocci, Marco et al. (May 2006). "Pain Activates Cortical Areas in the Preterm Newborn Brain". In: *PAIN* 122.1, pp. 109–117. ISSN: 0304-3959. DOI: 10.1016/j.pain.2006.01.015 (cited pages 32, 114).

Baxter, Luke et al. (May 2021). "Functional and Diffusion MRI Reveal the Neurophysiological Basis of Neonates' Noxious-Stimulus Evoked Brain Activity". In: *Nat Commun* 12.1, p. 2744. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22960-0 (cited page 33).

Baxter, Luke et al. (July 2023). "Study Protocol: Is the Measurement of Noxious Stimulus-Evoked EEG Activity a Reliable, Valid, and Interpretable Tool to Assess Analgesic Efficacy in Neonates?" In: DOI: 10.17605/OSF.IO/TPYBX (cited pages 25, 115).

Baxter, Luke et al. (June 2024). *Is Noxious Stimulus-Evoked Electroencephalography Response a Reliable, Valid, and Interpretable Outcome Measure to Assess Analgesic Efficacy in Neonates? A Systematic Review and Individual Participant Data (IPD) Meta-Analysis Protocol*. DOI: 10.21203/rs.3.rs-4464161/v1 (cited pages 79, 80).

Beer (1852). "Bestimmung Der Absorption Des Rothen Lichts in Farbigen Flüssigkeiten". In: *Annalen der Physik* 162.5, pp. 78–88. ISSN: 1521-3889. DOI: 10.1002/andp.18521620505 (cited page 32).

Belda, Sylvia et al. (2004). "Screening for Retinopathy of Prematurity: Is It Painful?" In: *Neonatology* 86.3, pp. 195–200. ISSN: 1661-7800, 1661-7819. DOI: 10.1159/000079542 (cited page 112).

Benedetti, Fabrizio et al. (Nov. 2005). "Neurobiological Mechanisms of the Placebo Effect". In: *J. Neurosci.* 25.45, pp. 10390–10402. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.3458-05.2005 (cited pages 12, 14).

Benoit, Britney et al. (Mar. 2021). "The Influence of Breastfeeding on Cortical and Bio-Behavioural Indicators of Procedural Pain in Newborns: Findings of a Randomized Controlled Trial". In: *Early Human Development* 154, p. 105308. ISSN: 0378-3782. DOI: 10.1016/j.earlhumdev.2021.105308 (cited pages 38, 68).

Berger, Hans (1929). "Über Das Elektroenkephalogramm Des Menschen". In: *Archiv für psychiatrie und nervenkrankheiten* 87.1, pp. 527–570 (cited page 19).

Bigdely-Shamlo, Nima et al. (2015). "The PREP Pipeline: Standardized Preprocessing for Large-Scale EEG Analysis". In: *Frontiers in Neuroinformatics* 9. ISSN: 1662-5196 (cited pages 120, 121).

Bland, Martin (2015). *An Introduction to Medical Statistics*. Oxford, UNITED KINGDOM: Oxford University Press, Incorporated. ISBN: 978-0-19-251839-2 (cited pages 46, 75).

Blencowe, Hannah et al. (June 2012). "National, Regional, and Worldwide Estimates of Preterm Birth Rates in the Year 2010 with Time Trends since 1990 for Selected Countries: A Systematic Analysis and Implications". In: *The Lancet* 379.9832, pp. 2162–2172. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(12)60820-4 (cited page 16).

Blum, Sarah et al. (2019). "A Riemannian Modification of Artifact Subspace Reconstruction for EEG Artifact Handling". In: *Frontiers in Human Neuroscience* 13. ISSN: 1662-5161 (cited page 59).

Bourel-Ponchel, Emilie et al. (Jan. 2021). "Normal EEG during the Neonatal Period: Maturational Aspects from Premature to Full-Term Newborns". In: *Neurophysiologie Clinique* 51.1, pp. 61–88. ISSN: 0987-7053. DOI: 10.1016/j.neucli.2020.10.004 (cited page 46).

Brahnam, Sheryl et al. (Mar. 2006). "Machine Recognition and Representation of Neonatal Facial Displays of Acute Pain". In: *Artificial Intelligence in Medicine* 36.3, pp. 211–222. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2004.12.003 (cited page 29).

Breiman, Leo (Oct. 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cited page 55).

Büchel, Christian et al. (Mar. 2014). "Placebo Analgesia: A Predictive Coding Perspective". In: *Neuron* 81.6, pp. 1223–1239. ISSN: 08966273. DOI: 10.1016/j.neuron.2014.02.042 (cited page 13).

Bucsea, Oana et al. (May 2023). "Clinical Thresholds in Pain-Related Facial Activity Linked to Differences in Cortical Network Activation in Neonates". In: *PAIN* 164.5, p. 1039. ISSN: 0304-3959. DOI: 10.1097/j.pain.0000000000002798 (cited pages 35, 112).

Bushnell, M. Catherine, Marta Čeko, and Lucie A. Low (July 2013). "Cognitive and Emotional Control of Pain and Its Disruption in Chronic Pain". In: *Nat Rev Neurosci* 14.7, pp. 502–511. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3516 (cited page 12).

Butruille, L. et al. (Mar. 2015). "Development of a Pain Monitoring Device Focused on Newborn Infant Applications: The NeoDoloris Project". In: *IRBM*. ANR TECSAN 2015 36.2, pp. 80–85. ISSN: 1959-0318. DOI: 10.1016/j.irbm.2015.01.005 (cited page 31).

Buzsaki, Gyorgy (2006). *Rhythms of the Brain*. Oxford university press (cited page 53).

Caldwell, C. D. and K. L. Watterberg (June 2015). "Effect of Premedication Regimen on Infant Pain and Stress Response to Endotracheal Intubation". In: *J Perinatol* 35.6, pp. 415–418. ISSN: 1476-5543. DOI: 10.1038/jp.2014.227 (cited page 112).

Carbajal, Ricardo et al. (July 2008). "Epidemiology and Treatment of Painful Procedures in Neonates in Intensive Care Units". In: *JAMA* 300.1, pp. 60–70. ISSN: 0098-7484. DOI: 10.1001/jama.300.1.60 (cited pages 6, 16, 108).

Castillo, Paul et al. (Mar. 2024). "EEG Response to a High Volume (1.5 mL/Kg) Caudal Block in Infants Less than 3 Months". In: *Reg Anesth Pain Med* 49.3, pp. 163–167. ISSN: 1098-7339, 1532-8651. DOI: 10.1136/rapm-2023-104452 (cited page 38).

Cerulli, Giovanni (2023). *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*. Statistics and Computing. Cham: Springer International Publishing. ISBN: 978-3-031-41336-0 978-3-031-41337-7. DOI: 10.1007/978-3-031-41337-7 (cited page 55).

Chang, Won-Du et al. (Feb. 2016). "Detection of Eye Blink Artifacts from Single Prefrontal Channel Electroencephalogram". In: *Computer Methods and Programs in Biomedicine* 124, pp. 19–30. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2015.10.011 (cited page 53).

Chauhan, Chhaviraj, Monika Agrawal, and Pooja Sabherwal (Dec. 2023). "Accurate QRS Complex Detection in 12-Lead ECG Signals Using Multi-Lead Fusion". In: *Measurement* 223, p. 113776. ISSN: 0263-2241. DOI: 10.1016/j.measurement.2023.113776 (cited page 106).

Chaumon, Maximilien, Dorothy V. M. Bishop, and Niko A. Busch (July 2015). "A Practical Guide to the Selection of Independent Components of the Electroencephalogram for Artifact Correction". In: *Journal of Neuroscience Methods*. Cutting-Edge EEG Methods 250, pp. 47–63. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2015.02.025 (cited page 48).

Cobo, Maria et al. (Apr. 2021). "Quantifying Noxious-Evoked Baseline Sensitivity in Neonates to Optimise Analgesic Trials". In: *eLife* 10, e65266. ISSN: 2050-084X. DOI: 10.7554/eLife.65266 (cited pages 24, 34, 40, 44–46, 85, 103).

Cobo, Maria et al. (July 2022). "Multicentre, Randomised Controlled Trial to Investigate the Effects of Parental Touch on Relieving Acute Procedural Pain in Neonates (Petal)". In: *BMJ Open* 12.7, e061841. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2022-061841 (cited pages 34, 40, 44–46, 51, 58, 68).

Colloca, Luana et al. (Feb. 2017). "Neuropathic Pain". In: *Nat Rev Dis Primers* 3.1, pp. 1–19. ISSN: 2056-676X. DOI: 10.1038/nrdp.2017.2 (cited page 12).

Committee on Fetus and Newborn (Nov. 2004). "Age Terminology During the Perinatal Period". In: *Pediatrics* 114.5, pp. 1362–1364. ISSN: 0031-4005. DOI: 10.1542/peds.2004-1915 (cited page 2).

Cook, Jonathan A. (May 2023). "C5Choosing the Sample Size for a Clinical Trial". In: *An Introduction to Clinical Trials*. Ed. by Jonathan A. Cook. Oxford University Press, p. 0. ISBN: 978-0-19-888523-8. DOI: 10.1093/med/9780198885238.003.0005 (cited page 75).

Cornelissen, Laura et al. (Oct. 2013). "Postnatal Temporal, Spatial and Modality Tuning of Nociceptive Cutaneous Flexion Reflexes in Human Infants". In: *PLoS ONE* 8.10. Ed. by Michael Costigan, e76470. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076470 (cited page 29).

Costa, Haline Tereza Matias De Lima et al. (Sept. 2018). "Use of Off-Label and Unlicensed Medicines in Neonatal Intensive Care". In: *PLoS ONE* 13.9. Ed. by Ricardo Queiroz Gurgel, e0204427. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0204427 (cited page 18).

Cowen, R. et al. (2015). "Assessing Pain Objectively: The Use of Physiological Markers". In: *Anaesthesia* 70.7, pp. 828–847. ISSN: 1365-2044. DOI: 10.1111/anae.13018 (cited pages 75, 106, 116).

Cowley, Benjamin U. and Jussi Korpela (2018). "Computational Testing for Automated Preprocessing 2: Practical Demonstration of a System for Scientific Data-Processing Workflow Management for High-Volume EEG". In: *Frontiers in Neuroscience* 12. ISSN: 1662-453X (cited page 122).

Cowley, Benjamin U., Jussi Korpela, and Jari Torniainen (Mar. 2017). "Computational Testing for Automated Preprocessing: A Matlab Toolbox to Enable Large Scale Electroencephalography Data Processing". In: *PeerJ Comput. Sci.* 3, e108. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.108 (cited pages 121, 122).

Craig, A (June 2003). "A New View of Pain as a Homeostatic Emotion". In: *Trends in Neurosciences* 26.6, pp. 303–307. ISSN: 01662236. DOI: 10.1016/S0166-2236(03)00123-1 (cited page 13).

Cruz, Fernandes, and Oliveira (2016). "Epidemiology of Painful Procedures Performed in Neonates: A Systematic Review of Observational Studies". In: *European Journal of Pain* 20.4, pp. 489–498. ISSN: 1532-2149. DOI: 10.1002/ejp.757 (cited pages 16, 108).

da Cruz, Janir Ramos et al. (July 2018). "An Automatic Pre-Processing Pipeline for EEG Analysis (APP) Based on Robust Statistics". In: *Clinical Neurophysiology* 129.7, pp. 1427–1437. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2018.04.600 (cited pages 121, 122).

Davidson, Megan (2014). "Known-Groups Validity". In: *Encyclopedia of Quality of Life and Well-Being Research*. Ed. by Alex C. Michalos. Dordrecht: Springer Netherlands, pp. 3481–3482. ISBN: 978-94-007-0752-8 978-94-007-0753-5. DOI: 10.1007/978-94-007-0753-5_1581 (cited page 77).

Davis, Karen D. et al. (Oct. 2017). "Brain Imaging Tests for Chronic Pain: Medical, Legal and Ethical Issues and Recommendations". In: *Nat Rev Neurol* 13.10, pp. 624–638. ISSN: 1759-4766. DOI: 10.1038/nrneurol.2017.122 (cited page 7).

Davis, Karen D. et al. (July 2020). "Discovery and Validation of Biomarkers to Aid the Development of Safe and Effective Pain Therapeutics: Challenges and Opportunities". In: *Nat Rev Neurol* 16.7, pp. 381–400. ISSN: 1759-4766. DOI: 10.1038/s41582-020-0362-2 (cited page 19).

de Vet, Henrica C. W. et al. (Aug. 2011). *Measurement in Medicine: A Practical Guide*. 1st ed. Cambridge University Press. ISBN: 978-0-521-13385-2 978-0-521-11820-0 978-0-511-99621-4. DOI: 10.1017/CBO9780511996214 (cited pages 26, 76, 77).

Debnath, Ranjan et al. (2020). "The Maryland Analysis of Developmental EEG (MADE) Pipeline". In: *Psychophysiology* 57.6, e13580. ISSN: 1469-8986. DOI: 10.1111/psyp.13580 (cited pages 49, 68, 109, 120–122, 139).

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson (1988). "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". In: *Biometrics* 44.3, pp. 837–845. ISSN: 0006-341X. DOI: 10.2307/2531595. JSTOR: 2531595 (cited page 94).

Delorme, Arnaud, Terrence Sejnowski, and Scott Makeig (Feb. 2007). "Enhanced Detection of Artifacts in EEG Data Using Higher-Order Statistics and Independent Component Analysis". In: *NeuroImage* 34.4, pp. 1443–1449. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2006.11.004 (cited pages 49, 70).

Desjardins, James A. et al. (Jan. 2021). "EEG Integrated Platform Lossless (EEG-IP-L) Pre-Processing Pipeline for Objective Signal Quality Assessment Incorporating Data Annotation and Blind Source Separation". In: *Journal of Neuroscience Methods* 347, p. 108961. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2020.108961 (cited pages 121, 122).

Dogan, Hulya and Ramazan Ozgur Dogan (July 2023). "A Comprehensive Review of Computer-based Techniques for R-Peaks/QRS Complex Detection in ECG Signal". In: *Arch Computat Methods Eng* 30.6, pp. 3703–3721. ISSN: 1886-1784. DOI: 10.1007/s11831-023-09916-x (cited page 106).

Donoghue, Thomas et al. (Dec. 2020). "Parameterizing Neural Power Spectra into Periodic and Aperiodic Components". In: *Nat Neurosci* 23.12, pp. 1655–1665. ISSN: 1097-6256. DOI: 10.1038/s41593-020-00744-x (cited page 54).

Duff, Eugene P. et al. (Sept. 2020). "Inferring Pain Experience in Infants Using Quantitative Whole-Brain Functional MRI Signatures: A Cross-Sectional, Observational Study". In: *The Lancet Digital Health* 2.9, e458–e467. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(20)30168-0 (cited pages 17, 32, 33, 114).

Elm, Erik von et al. (Oct. 2007). "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies". In: *The Lancet* 370.9596, pp. 1453–1457. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(07)61602-X (cited pages 144–146).

Fabrizi, Lorenzo et al. (Sept. 2011). "A Shift in Sensory Processing That Enables the Developing Human Brain to Discriminate Touch from Pain". In: *Current Biology* 21.18, pp. 1552–1558. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.08.010 (cited pages 68, 91).

Fabrizi, Lorenzo et al. (June 2016). "Encoding of Mechanical Nociception Differs in the Adult and Infant Brain". In: *Sci Rep* 6.1, p. 28642. ISSN: 2045-2322. DOI: 10.1038/srep28642 (cited page 17).

FDA (2009). *Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling ClaimsGuidance for Industry*. Tech. rep. (cited page 76).

FDA-NIH Biomarker Working Group (2016). *BEST (Biomarkers, EndpointS, and Other Tools) Resource*. Silver Spring (MD): Food and Drug Administration (US) (cited page 25).

— (Nov. 2021). "Glossary". In: *BEST (Biomarkers, EndpointS, and Other Tools) Resource [Internet]*. Food and Drug Administration (US) (cited pages 24, 77).

Fields, Howard L. (Sept. 2018). "How Expectations Influence Pain". In: *PAIN* 159, S3. ISSN: 0304-3959. DOI: 10.1097/j.pain.0000000000001272 (cited page 13).

Fitzgerald, Maria (Dec. 2015). "What Do We Really Know about Newborn Infant Pain?" In: *Experimental Physiology* 100.12, pp. 1451–1457. ISSN: 0958-0670, 1469-445X. DOI: 10.1113/EP085134 (cited pages 28, 30).

Fitzgerald, Maria and Suellen M Walker (Jan. 2009). "Infant Pain Management: A Developmental Neurobiological Approach". In: *Nat Rev Neurol* 5.1, pp. 35–50. ISSN: 1745-834X, 1745-8358. DOI: 10.1038/ncpneuro0984 (cited pages 31, 114).

Fleiss, Joseph L. (1981). *Statistical Methods for Rates and Proportions. Second Edition*. Wiley, John and Sons, Incorporated, New York, N.Y. (cited page 59).

Flint, Robert B. et al. (June 2018). "Large Differences in Neonatal Drug Use between NICUs Are Common Practice: Time for Consensus?" In: *Brit J Clinical Pharma* 84.6, pp. 1313–1323. ISSN: 0306-5251, 1365-2125. DOI: 10.1111/bcp.13563 (cited page 19).

Fló, Ana et al. (Apr. 2022). "Automated Pipeline for Infants Continuous EEG (APICE): A Flexible Pipeline for Developmental Cognitive Studies". In: *Developmental Cognitive Neuroscience* 54, p. 101077. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2022.101077 (cited pages 49, 68, 109, 121, 122).

Foster, Jann P, Christine Taylor, and Kaye Spence (Feb. 2017). "Topical Anaesthesia for Needle-Related Pain in Newborn Infants". In: *Cochrane Database of Systematic Reviews* 2017.2. Ed. by Cochrane Neonatal Group. ISSN: 14651858. DOI: 10.1002/14651858.CD010331.pub2 (cited page 18).

Franck, Linda Sturla (1986). "A New Method To Quantitatively Describe Pain Behavior in Infants". In: *Nursing Research* 35.1, p. 28. ISSN: 0029-6562 (cited pages 28, 113).

Fujioka, Takako et al. (Jan. 2011). "Comparison of Artifact Correction Methods for Infant EEG Applied to Extraction of Event-Related Potential Signals". In: *Clinical Neurophysiology* 122.1, pp. 43–51. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2010.04.036 (cited pages 36, 48).

Fuller, Barbara F. (1991). "Acoustic Discrimination of Three Types of Infant Cries". In: *Nursing Research* 40.3, p. 156. ISSN: 0029-6562 (cited pages 28, 113).

Gabard-Durnam, Laurel J. et al. (2018). "The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data". In: *Frontiers in Neuroscience* 12. ISSN: 1662-453X (cited pages 49, 109, 121, 122, 139).

Garcí-Larrea, Luis et al. (Dec. 1997). "Association and Dissociation between Laser-Evoked Potentials and Pain Perception". In: *NeuroReport* 8.17, p. 3785. ISSN: 0959-4965 (cited page 34).

Gee, Alan H. et al. (Sept. 2017). "Predicting Bradycardia in Preterm Infants Using Point Process Analysis of Heart Rate". In: *IEEE Trans. Biomed. Eng.* 64.9, pp. 2300–2308. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2016.2632746 (cited page 31).

Geißler, Carmen et al. (Sept. 2020). "Drug Utilisation and Off-Label Use on a German Neonatal Intensive Care Unit: A Retrospective Cohort Study and 10-Year Comparison". In: *Pharmacy* 8.3, p. 173. ISSN: 2226-4787. DOI: 10.3390/pharmacy8030173 (cited page 18).

Gendras, Julie et al. (May 2021). "The Newborn Infant Parasympathetic Evaluation Index for Acute Procedural Pain Assessment in Preterm Infants". In: *Pediatr Res* 89.7, pp. 1840–1847. ISSN: 1530-0447. DOI: 10.1038/s41390-020-01152-4 (cited page 31).

Georgieva, Stanimira et al. (2020). "Toward the Understanding of Topographical and Spectral Signatures of Infant Movement Artifacts in Naturalistic EEG". In: *Frontiers in Neuroscience* 14. ISSN: 1662-453X (cited page 48).

Ghandeharion, Hosna and Abbas Erfanian (Sept. 2010). "A Fully Automatic Ocular Artifact Suppression from EEG Data Using Higher Order Statistics: Improved Performance by Wavelet Analysis". In: *Medical Engineering & Physics* 32.7, pp. 720–729. ISSN: 13504533. DOI: 10.1016/j.medengphy.2010.04.010 (cited page 54).

Gibbins, Sharyn et al. (2008). "Comparison of Pain Responses in Infants of Different Gestational Ages". In: *Neonatology* 93.1, pp. 10–18. ISSN: 1661-7800, 1661-7819. DOI: 10.1159/000105520 (cited pages 28, 29, 31).

Gibbins, Sharyn et al. (Apr. 2014). "Validation of the Premature Infant Pain Profile-Revised (PIPP-R)". In: *Early Hum Dev* 90.4, pp. 189–193. ISSN: 1872-6232. DOI: 10.1016/j.earlhumdev.2014.01.005 (cited pages 7, 28).

Giesecke, Thorsten et al. (2005). "The Relationship between Depression, Clinical Pain, and Experimental Pain in a Chronic Pain Cohort". In: *Arthritis & Rheumatism* 52.5, pp. 1577–1584. ISSN: 1529-0131. DOI: 10.1002/art.21008 (cited page 14).

Giordano, Vito et al. (Dec. 2019). "Pain and Sedation Scales for Neonatal and Pediatric Patients in a Preverbal Stage of Development: A Systematic Review". In: *JAMA Pediatr* 173.12, p. 1186. ISSN: 2168-6203. DOI: 10.1001/jamapediatrics.2019.3351 (cited page 28).

Giordano, Vito et al. (Aug. 2023). "Pain and Neurodevelopmental Outcomes of Infants Born Very Preterm". In: *Develop Med Child Neuro* 65.8, pp. 1043–1052. ISSN: 0012-1622, 1469-8749. DOI: 10.1111/dmcn.15505 (cited pages 17, 110, 111).

Glenzel, Luana et al. (Oct. 2023). "Validity and Reliability of Pain and Behavioral Scales for Preterm Infants: A Systematic Review". In: *Pain Management Nursing* 24.5, e84–e96. ISSN: 1524-9042. DOI: 10.1016/j.pmn.2023.06.010 (cited pages 7, 28, 79, 106).

Goksan, Sezgi et al. (Apr. 2015). "fMRI Reveals Neural Activity Overlap between Adult and Infant Pain". In: *eLife* 4. Ed. by Jody C Culham, e06356. ISSN: 2050-084X. DOI: 10.7554/eLife.06356 (cited pages 16, 17).

Goksan, Sezgi et al. (Sept. 2018). "The Influence of the Descending Pain Modulatory System on Infant Pain-Related Brain Activity". In: *eLife* 7. Ed. by Peggy Mason and Eve Marder, e37125. ISSN: 2050-084X. DOI: 10.7554/eLife.37125 (cited page 15).

González-Villar, Alberto J. et al. (Dec. 2020). "Patients with Fibromyalgia Show Increased Beta Connectivity across Distant Networks and Microstates Alterations in Resting-State Electroencephalogram". In: *NeuroImage* 223, p. 117266. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117266 (cited page 35).

Green, Gabrielle et al. (Feb. 2019). "Behavioural Discrimination of Noxious Stimuli in Infants Is Dependent on Brain Maturation". In: *Pain* 160.2, pp. 493–500. ISSN:

0304-3959, 1872-6623. DOI: 10.1097/j.pain.0000000000001425 (cited pages 29, 30, 75, 120).

Grunau, Ruth E. et al. (May 2009). "Neonatal Pain, Parenting Stress and Interaction, in Relation to Cognitive and Motor Development at 8 and 18 Months in Preterm Infants". In: *PAIN* 143.1, pp. 138–146. ISSN: 0304-3959. DOI: 10.1016/j.pain.2009.02.014 (cited page 17).

Gursul, Deniz et al. (Dec. 2018). "Stroking Modulates Noxious-Evoked Brain Activity in Human Infants". In: *Current Biology* 28.24, R1380–R1381. ISSN: 09609822. DOI: 10.1016/j.cub.2018.11.014 (cited page 68).

Haar, Alfred (Sept. 1910). "Zur Theorie der orthogonalen Funktionensysteme". In: *Math. Ann.* 69.3, pp. 331–371. ISSN: 1432-1807. DOI: 10.1007/BF01456326 (cited pages 54, 113).

Hari, Riitta and Aina Puce (Sept. 2023). *MEG - EEG Primer*. 2nd ed. Oxford University PressNew York. ISBN: 978-0-19-754218-7 978-0-19-754221-7. DOI: 10.1093/med/9780197542187.001.0001 (cited page 33).

"Must We Suffer in Order to Stay Healthy?" (Aug. 2018). "Must We Suffer in Order to Stay Healthy? Pleasure and Pain in Ancient Medical Literature". In: *Pain and Pleasure in Classical Times*. Ed. by William V. Harris. BRILL, pp. 36–54. ISBN: 978-90-04-37950-3 978-90-04-37949-7. DOI: 10.1163/9789004379503_004 (cited page 12).

Hartley, Caroline and Rebeccah Slater (2014). "Neurophysiological Measures of Nociceptive Brain Activity in the Newborn Infant – the next Steps". In: *Acta Paediatrica* 103.3, pp. 238–242. ISSN: 1651-2227. DOI: 10.1111/apa.12490 (cited pages 76, 106, 116).

Hartley, Caroline et al. (July 2015). "The Relationship between Nociceptive Brain Activity, Spinal Reflex Withdrawal and Behaviour in Newborn Infants". In: *Scientific Reports* 5.1, p. 12519. ISSN: 2045-2322. DOI: 10.1038/srep12519 (cited pages 23, 34, 37, 40, 41, 44, 45, 48, 49, 76, 87, 116, 139).

Hartley, Caroline et al. (May 2017). "Nociceptive Brain Activity as a Measure of Analgesic Efficacy in Infants". In: *Sci. Transl. Med.* 9.388, eaah6122. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.aah6122 (cited pages 8, 23–25, 37, 40, 44, 45, 76, 81, 85, 86, 91, 103, 105, 111, 116, 126, 139).

Hartley, Caroline et al. (Dec. 2018). "Analgesic Efficacy and Safety of Morphine in the Procedural Pain in Premature Infants (Poppi) Study: Randomised Placebo-Controlled Trial". In: *The Lancet* 392.10164, pp. 2595–2605. ISSN: 01406736. DOI: 10.1016/S0140-6736(18)31813-0 (cited pages 18, 68, 111, 120).

Hatfield, Linda A. and Elizabeth A. Ely (Jan. 2015). "Measurement of Acute Pain in Infants: A Review of Behavioral and Physiological Variables". In: *Biological Research For Nursing* 17.1, pp. 100–111. ISSN: 1099-8004, 1552-4175. DOI: 10.1177/1099800414531448 (cited pages 28–30, 113).

Hatz, F. et al. (Feb. 2015). "Reliability of Fully Automated versus Visually Controlled Pre- and Post-Processing of Resting-State EEG". In: *Clinical Neurophysiology* 126.2, pp. 268–274. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2014.05.014 (cited page 48).

Hauck, Annalisa G. V. et al. (Apr. 2024). "Effect of Parental Touch on Relieving Acute Procedural Pain in Neonates and Parental Anxiety (Petal): A Multicentre,

Randomised Controlled Trial in the UK". In: *The Lancet Child & Adolescent Health* 8.4, pp. 259–269. ISSN: 2352-4642, 2352-4650. DOI: 10.1016/S2352-4642(23)00340-1 (cited pages 42, 85, 103, 123, 140).

Heiderich, Tatiany M. et al. (Nov. 2023). "Face-Based Automatic Pain Assessment: Challenges and Perspectives in Neonatal Intensive Care Units". In: *Jornal de Pediatria* 99.6, pp. 546–560. ISSN: 0021-7557. DOI: 10.1016/j.jped.2023.05.005 (cited page 30).

Heiderich, Tatiany Marcondes, Ana Teresa Figueiredo Stochero Leslie, and Ruth Guinsburg (2015). "Neonatal Procedural Pain Can Be Assessed by Computer Software That Has Good Sensitivity and Specificity to Detect Facial Movements". In: *Acta Paediatrica* 104.2, e63–e69. ISSN: 1651-2227. DOI: 10.1111/apa.12861 (cited pages 29, 114).

Hermann, Christiane et al. (Dec. 2006). "Long-Term Alteration of Pain Sensitivity in School-Aged Children with Early Pain Experiences". In: *Pain* 125.3, pp. 278–285. ISSN: 0304-3959. DOI: 10.1016/j.pain.2006.08.026 (cited page 18).

Hermans, Tim et al. (Feb. 2023). "A Multi-Task and Multi-Channel Convolutional Neural Network for Semi-Supervised Neonatal Artefact Detection". In: *J. Neural Eng.* ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/acbc4b (cited page 49).

Higgins, Julian Pt et al. (Sept. 2019). "Assessing Risk of Bias in a Randomized Trial". In: *Cochrane Handbook for Systematic Reviews of Interventions*. Ed. by Julian P.T. Higgins et al. 1st ed. Wiley, pp. 205–228. ISBN: 978-1-119-53662-8 978-1-119-53660-4. DOI: 10.1002/9781119536604.ch8 (cited page 19).

Higuchi, T. (June 1988). "Approach to an Irregular Time Series on the Basis of the Fractal Theory". In: *Physica D: Nonlinear Phenomena* 31.2, pp. 277–283. ISSN: 01672789. DOI: 10.1016/0167-2789(88)90081-4 (cited pages 54, 113).

Hock, Christoph et al. (1997). "Decrease in Parietal Cerebral Hemoglobin Oxygenation during Performance of a Verbal Fluency Task in Patients with Alzheimer's Disease Monitored by Means of near-Infrared Spectroscopy Ž NIRS. — Correlation with Simultaneous rCBF-PET Measurements". In: (cited page 32).

Holsti, Liisa et al. (Nov. 2005). "Body Movements: An Important Additional Factor in Discriminating Pain From Stress in Preterm Infants". In: *The Clinical Journal of Pain* 21.6, pp. 491–498. ISSN: 0749-8047. DOI: 10.1097/01.ajp.0000146163.30776.44 (cited pages 29, 31, 113, 114).

Iannetti, G. D. et al. (Jan. 2005). "Operculoinsular Cortex Encodes Pain Intensity at the Earliest Stages of Cortical Processing as Indicated by Amplitude of Laser-Evoked Potentials in Humans". In: *Neuroscience* 131.1, pp. 199–208. ISSN: 0306-4522. DOI: 10.1016/j.neuroscience.2004.10.035 (cited page 34).

Iannetti, G. D. et al. (Sept. 2013). "Pinprick-Evoked Brain Potentials: A Novel Tool to Assess Central Sensitization of Nociceptive Pathways in Humans". In: *Journal of Neurophysiology* 110.5, pp. 1107–1116. ISSN: 0022-3077. DOI: 10.1152/jn.00774.2012 (cited page 34).

International Association for the Study of Pain, ed. (1994). *Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms*. 2. ed. Seattle: IASP Press. ISBN: 978-0-931092-05-3 (cited page 13).

Intriligator, James and John Polich (June 1995). "On the Relationship between EEG and ERP Variability". In: *International Journal of Psychophysiology* 20.1, pp. 59–74. ISSN: 0167-8760. DOI: 10.1016/0167-8760(95)00028-Q (cited page 105).

ISO 13485 (2016). *Medical Devices. Quality Management Systems.* DOI: 10.3403/30257645 (cited page 114).

Jasper, Herbert (May 1958). "Report of the Committee on Methods of Clinical Examination in Electroencephalography: 1957". In: *Electroencephalography and Clinical Neurophysiology* 10.2, pp. 370–375. ISSN: 0013-4694. DOI: 10.1016/0013-4694(58)90053-1 (cited page 21).

Jenkinson, Mark and Michael Chappell (2018). *Introduction to Neuroimaging Analysis.* First edition. Oxford Neuroimaging Primers. New York, NY: Oxford University Press. ISBN: 978-0-19-881630-0 (cited page 32).

Johnston, C. Celeste et al. (June 1995). "Differential Response to Pain by Very Premature Neonates". In: *Pain* 61.3, pp. 471–479. ISSN: 0304-3959. DOI: 10.1016/0304-3959(94)00213-X (cited page 30).

Jones, Laura et al. (Dec. 2017). "Nociceptive Cortical Activity Is Dissociated from Nociceptive Behavior in Newborn Human Infants under Stress". In: *Current Biology* 27.24, 3846–3851.e3. ISSN: 09609822. DOI: 10.1016/j.cub.2017.10.063 (cited page 30).

Jones, Laura et al. (Nov. 2018). "EEG, Behavioural and Physiological Recordings Following a Painful Procedure in Human Neonates". In: *Sci Data* 5.1, p. 180248. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.248 (cited pages 38, 46).

Julius, David and Allan I Basbaum (2001). "Molecular Mechanisms of Nociception". In: 413 (cited page 14).

Jung, Tzyy-Ping et al. (Mar. 2000). "Removing Electroencephalographic Artifacts by Blind Source Separation". In: *Psychophysiology* 37.2, pp. 163–178. ISSN: 1469-8986, 0048-5772. DOI: 10.1111/1469-8986.3720163 (cited page 49).

Kahnt, Thorsten et al. (Mar. 2010). "The Neural Code of Reward Anticipation in Human Orbitofrontal Cortex". In: *Proceedings of the National Academy of Sciences* 107.13, pp. 6010–6015. DOI: 10.1073/pnas.0912838107 (cited page 16).

Katz, Michael J. (Jan. 1988). "Fractals and the Analysis of Waveforms". In: *Computers in Biology and Medicine* 18.3, pp. 145–156. ISSN: 00104825. DOI: 10.1016/0010-4825(88)90041-8 (cited pages 54, 113).

Keil, Andreas et al. (2014). "Committee Report: Publication Guidelines and Recommendations for Studies Using Electroencephalography and Magnetoencephalography". In: *Psychophysiology* 51.1, pp. 1–21. ISSN: 1469-8986. DOI: 10.1111/psyp.12147 (cited page 139).

Kelleher, John D., Brian MacNamee, and Aoife D'Arcy (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* Cambridge, Massachusetts London, England: The MIT Press. ISBN: 978-0-262-02944-5 (cited page 56).

Kennedy-Shaffer, Lee (Mar. 2019). "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing". In: *The American Statistician* 73.sup1, pp. 82–90. ISSN: 0003-1305. DOI: 10.1080/00031305.2018.1537891 (cited page 104).

Khanna, Arjun et al. (Feb. 2015). "Microstates in Resting-State EEG: Current Status and Future Directions". In: *Neuroscience & Biobehavioral Reviews* 49, pp. 105–113. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2014.12.010 (cited page 35).

Kleinschmidt, Andreas et al. (Sept. 1996). "Simultaneous Recording of Cerebral Blood Oxygenation Changes during Human Brain Activation by Magnetic Resonance Imaging and Near-Infrared Spectroscopy". In: *J Cereb Blood Flow Metab* 16.5, pp. 817–826. ISSN: 0271-678X, 1559-7016. DOI: 10.1097/00004647-199609000-00006 (cited page 32).

Knaepen, Liesbeth et al. (2013). "Neonatal Repetitive Needle Pricking: Plasticity of the Spinal Nociceptive Circuit and Extended Postoperative Pain in Later Life". In: *Developmental Neurobiology* 73.1, pp. 85–97. ISSN: 1932-846X. DOI: 10.1002/dneu.22047 (cited page 18).

Kragel, Philip A. et al. (July 2018). "Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging". In: *Neuron* 99.2, pp. 257–273. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.06.009 (cited page 33).

Kumaravel, Velu Prabhakar et al. (Nov. 2021). "Efficient Artifact Removal from Low-Density Wearable EEG Using Artifacts Subspace Reconstruction". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 333–336. DOI: 10.1109/EMBC46164.2021.9629771 (cited page 49).

Kumaravel, Velu Prabhakar et al. (Apr. 2022). "NEAR: An Artifact Removal Pipeline for Human Newborn EEG Data". In: *Developmental Cognitive Neuroscience* 54, p. 101068. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2022.101068 (cited pages 49, 68, 109, 121).

Lachaux, Jean-Philippe et al. (1999). "Measuring Phase Synchrony in Brain Signals". In: *Human Brain Mapping* 8.4, pp. 194–208. ISSN: 1097-0193. DOI: 10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C (cited page 22).

"Concise Medical Dictionary" (Feb. 2020). In: *Concise Medical Dictionary*. Ed. by Jonathan Law and Elizabeth Martin. Oxford University Press. ISBN: 978-0-19-883661-2 (cited page 2).

Lazzaro, I. et al. (Jan. 2001). "The Modulation of Late Component Event Related Potentials by Pre-Stimulus EEG Theta Activity in ADHD". In: *International Journal of Neuroscience* 107.3-4, pp. 247–264. ISSN: 0020-7454. DOI: 10.3109/00207450109150688 (cited page 105).

Liu, Feifei et al. (2018). "Performance Analysis of Ten Common QRS Detectors on Different ECG Application Cases". In: *Journal of Healthcare Engineering* 2018.1, p. 9050812. ISSN: 2040-2309. DOI: 10.1155/2018/9050812 (cited page 106).

Llerena, Amelia et al. (Feb. 2023). "Neonatal Pain Assessment: Do We Have the Right Tools?" In: *Front. Pediatr.* 10. ISSN: 2296-2360. DOI: 10.3389/fped.2022.1022751 (cited page 79).

Lopes da Silva, Fernando (Dec. 2013). "EEG and MEG: Relevance to Neuroscience". In: *Neuron* 80.5, pp. 1112–1128. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2013.10.017 (cited pages 20, 21).

Luck, Steven J. (Steven John) (2014). *An Introduction to the Event-Related Potential Technique*. Second edition. Cambridge, Massachusetts: The MIT Press. ISBN: 0-262-32406-7 (cited pages 21–23, 35, 36, 48, 49, 51, 53, 84, 132).

Maimon, Neta et al. (Dec. 2013). "Electroencephalographic Activity in Response to Procedural Pain in Preterm Infants Born at 28 and 33 Weeks Gestational Age". In: *The Clinical Journal of Pain* 29.12, p. 1044. ISSN: 0749-8047. DOI: 10.1097/AJP.0b013e318284e525 (cited pages 38, 68).

Maitre, Nathalie L et al. (Sept. 2017). "Cry Presence and Amplitude Do Not Reflect Cortical Processing of Painful Stimuli in Newborns with Distinct Responses to Touch or Cold". In: *Arch Dis Child Fetal Neonatal Ed* 102.5, F428–F433. ISSN: 1359-2998, 1468-2052. DOI: 10.1136/archdischild-2016-312279 (cited page 38).

Makeig, Scott et al. (1995). "Independent Component Analysis of Electroencephalographic Data". In: *Advances in Neural Information Processing Systems 8* (cited page 49).

Marchant, Simon et al. (Aug. 2024). "A Machine Learning Artefact Detection Method for Single-Channel Infant Event-Related Potential Studies". In: *J. Neural Eng.* 21.4, p. 046021. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ad5c04 (cited pages 48, 109).

Maresova, Petra et al. (Oct. 2015). "The Potential of Medical Device Industry in Technological and Economical Context". In: *Therapeutics and Clinical Risk Management* 11, pp. 1505–1514. ISSN: 1176-6336. DOI: 10.2147/TCRM.S88574 (cited page 114).

Marquand, Andre et al. (Feb. 2010). "Quantitative Prediction of Subjective Pain Intensity from Whole-Brain fMRI Data Using Gaussian Processes". In: *NeuroImage* 49.3, pp. 2178–2189. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2009.10.072 (cited page 33).

McLoughlin, Gráinne, Scott Makeig, and Ming T. Tsuang (2014). "In Search of Biomarkers in Psychiatry: EEG-based Measures of Brain Function". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 165.2, pp. 111–121. ISSN: 1552-485X. DOI: 10.1002/ajmg.b.32208 (cited page 22).

McNemar, Quinn (June 1947). "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages". In: *Psychometrika* 12.2, pp. 153–157. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02295996 (cited pages 55, 58).

McShane, Blakeley B. et al. (Mar. 2019). "Abandon Statistical Significance". In: *The American Statistician* 73.sup1, pp. 235–245. ISSN: 0003-1305. DOI: 10.1080/00031305.2018.1527253 (cited page 104).

Melzack, Ronald and Kenneth L Casey (1968). "Sensory, Motivational, and Central Control Determinants of Pain: A New Conceptual Model". In: *The skin senses* 1, pp. 423–43 (cited page 13).

Michel, Christoph M. and Thomas Koenig (Oct. 2018). "EEG Microstates as a Tool for Studying the Temporal Dynamics of Whole-Brain Neuronal Networks: A Review". In: *NeuroImage*. Brain Connectivity Dynamics 180, pp. 577–593. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.11.062 (cited page 35).

Mognon, Andrea et al. (2011). "ADJUST: An Automatic EEG Artifact Detector Based on the Joint Use of Spatial and Temporal Features". In: *Psychophysiology*

48.2, pp. 229–240. ISSN: 1469-8986. DOI: 10.1111/j.1469-8986.2010.01061.x (cited pages 48, 53).

Mokkink, Lidwine B et al. (July 2023). "Studies on Reliability and Measurement Error of Measurements in Medicine – From Design to Statistics Explained for Medical Researchers". In: *PROM* Volume 14, pp. 193–212. ISSN: 1179-271X. DOI: 10.2147/PROM.S398886 (cited page 77).

Mokkink, Lidwine B. et al. (July 2010). "The COSMIN Study Reached International Consensus on Taxonomy, Terminology, and Definitions of Measurement Properties for Health-Related Patient-Reported Outcomes". In: *Journal of Clinical Epidemiology* 63.7, pp. 737–745. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2010.02.006 (cited pages 26, 76–78, 105).

Molfese, Dennis L. (May 2000). "Predicting Dyslexia at 8 Years of Age Using Neonatal Brain Responses". In: *Brain and Language* 72.3, pp. 238–245. ISSN: 0093-934X. DOI: 10.1006/brln.2000.2287 (cited pages 49, 116).

Mörelius, Evalotte et al. (Oct. 2006). "Is a Nappy Change Stressful to Neonates?" In: *Early Human Development* 82.10, pp. 669–676. ISSN: 03783782. DOI: 10.1016/j.earlhumdev.2005.12.013 (cited page 29).

Mullen, Tim R. et al. (Nov. 2015). "Real-Time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG". In: *IEEE Trans Biomed Eng* 62.11, pp. 2553–2567. ISSN: 0018-9294. DOI: 10.1109/TBME.2015.2481482 (cited pages 48, 49).

Munsters, Josanne et al. (Jan. 2012). "Skin Conductance Measurements as Pain Assessment in Newborn Infants Born at 22–27weeks Gestational Age at Different Postnatal Age". In: *Early Human Development* 88.1, pp. 21–26. ISSN: 0378-3782. DOI: 10.1016/j.earlhumdev.2011.06.010 (cited pages 31, 114).

Nolan, H., R. Whelan, and R. B. Reilly (Sept. 2010). "FASTER: Fully Automated Statistical Thresholding for EEG Artifact Rejection". In: *Journal of Neuroscience Methods* 192.1, pp. 152–162. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2010.07.015 (cited pages 48, 53).

Norman, Elisabeth et al. (Oct. 2008). "Electroencephalographic Response to Procedural Pain in Healthy Term Newborn Infants". In: *Pediatr Res* 64.4, pp. 429–434. ISSN: 1530-0447. DOI: 10.1203/PDR.0b013e3181825487 (cited pages 38, 68).

Olejniczak, Piotr (June 2006). "Neurophysiologic Basis of EEG". In: *Journal of Clinical Neurophysiology* 23.3, p. 186. ISSN: 0736-0258. DOI: 10.1097/01.wnp.0000220079.61973.6c (cited page 20).

Orhan, Umut, Mahmut Hekim, and Mahmut Ozer (Sept. 2011). "EEG Signals Classification Using the K-means Clustering and a Multilayer Perceptron Neural Network Model". In: *Expert Systems with Applications* 38.10, pp. 13475–13481. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2011.04.149 (cited pages 54, 113).

Ougham, K and N Modi (Mar. 2017). *The NDAU Report 2016*. Tech. rep. Imperial College London. DOI: 10.25561/87363 (cited page 16).

Pavel, Andreea M et al. (June 2019). "GP252 The Effect of Music Therapy on the Electroencephalogram (EEG) and Heart Rate Variability (HRV) of Premature Infants during Routine Painful Procedures". In: *Abstracts*. BMJ Publishing Group Ltd and Royal College of Paediatrics and Child Health, A135.1–A135. DOI: 10.1136/archdischild-2019-epa.311 (cited page 38).

Pedregosa, Fabian et al. (Nov. 2011). "Scikit-Learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12.null, pp. 2825–2830. ISSN: 1532-4435 (cited page 55).

Pedroni, Andreas, Amirreza Bahreini, and Nicolas Langer (Oct. 2019). "Automagic: Standardized Preprocessing of Big EEG Data". In: *NeuroImage* 200, pp. 460–473. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.06.046 (cited pages 121, 122).

Pernet, Cyril R. et al. (June 2019). "EEG-BIDS, an Extension to the Brain Imaging Data Structure for Electroencephalography". In: *Sci Data* 6.1, p. 103. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0104-8 (cited page 124).

Pillai Riddell, Rebecca R. and Kenneth D. Craig (June 2007). "Judgments of Infant Pain: The Impact of Caregiver Identity and Infant Age". In: *Journal of Pediatric Psychology* 32.5, pp. 501–511. ISSN: 0146-8693. DOI: 10.1093/jpepsy/jsl049 (cited page 29).

Pion-Tonachini, Luca, Ken Kreutz-Delgado, and Scott Makeig (Sept. 2019). "ICLabel: An Automated Electroencephalographic Independent Component Classifier, Dataset, and Website". In: *NeuroImage* 198, pp. 181–197. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.05.026 (cited pages 49, 69).

Poldrack, Russell A. (Feb. 2006). "Can Cognitive Processes Be Inferred from Neuroimaging Data?" In: *Trends in Cognitive Sciences* 10.2, pp. 59–63. ISSN: 1364-6613. DOI: 10.1016/j.tics.2005.12.004 (cited page 33).

Pomerleau-Malcuit, Andrée and Rachel K. Clifton (1973). "Neonatal Heart-Rate Response to Tactile, Auditory, and Vestibular Stimulation in Different States". In: *Child Development* 44.3, pp. 485–496. ISSN: 0009-3920. DOI: 10.2307/1128003. JSTOR: 1128003 (cited page 91).

Pontifex, Matt (n.d.). *Loadcurry* (cited page 122).

Price, Donald D. and Ronald Dubner (Aug. 1977). "Neurons That Subserve the Sensory-Discriminative Aspects of Pain". In: *Pain* 3.4, pp. 307–338. ISSN: 0304-3959. DOI: 10.1016/0304-3959(77)90063-X (cited page 14).

Prinsen, C. A. C. et al. (May 2018). "COSMIN Guideline for Systematic Reviews of Patient-Reported Outcome Measures". In: *Qual Life Res* 27.5, pp. 1147–1157. ISSN: 0962-9343, 1573-2649. DOI: 10.1007/s11136-018-1798-3 (cited page 37).

Raja, Srinivasa N. et al. (Sept. 2020). "The Revised International Association for the Study of Pain Definition of Pain: Concepts, Challenges, and Compromises". In: *PAIN* 161.9, pp. 1976–1982. ISSN: 0304-3959. DOI: 10.1097/j.pain.0000000000001939 (cited pages 7, 12, 28, 30, 80).

Ranger, Manon, C. Céleste Johnston, and K.J.S. Anand (Oct. 2007). "Current Controversies Regarding Pain Assessment in Neonates". In: *Seminars in Perinatology* 31.5, pp. 283–288. ISSN: 01460005. DOI: 10.1053/j.semperi.2007.07.003 (cited pages 29–31).

Ranger, Manon et al. (Aug. 2015). "Neonatal Pain and Infection Relate to Smaller Cerebellum in Very Preterm Children at School Age". In: *The Journal of Pediatrics* 167.2, 292–298.e1. ISSN: 0022-3476. DOI: 10.1016/j.jpeds.2015.04.055 (cited pages 6, 18, 110).

Recher, Morgan et al. (Oct. 2021). "The Newborn Infant Parasympathetic Evaluation in Pediatric and Neonatology: A Literature Review". In: *J Clin Monit Comput*

35.5, pp. 959–966. ISSN: 1573-2614. DOI: 10.1007/s10877-021-00670-8 (cited page 31).

Regulation (EU) 2017/745 of the European Parliament (Apr. 2017) (cited page 115).

Rejer, I. and P. Gorski (2015). "Benefits of ICA in the Case of a Few Channel EEG". In: *Annu Int Conf IEEE Eng Med Biol Soc* 2015, pp. 7434–7437. ISSN: 2694-0604. DOI: 10.1109/EMBC.2015.7320110 (cited pages 36, 49).

Relland, Lance M. et al. (Mar. 2024). "Vibration-Based Mitigation of Noxious-Evoked Responses to Skin Puncture in Neonates and Infants: A Randomised Controlled Trial". In: *Arch Dis Child Fetal Neonatal Ed*, fetalneonatal-2023–326588. ISSN: 1359-2998, 1468-2052. DOI: 10.1136/archdischild-2023-326588 (cited page 38).

Rodkey, Elissa N. and Rebecca Riddell (Apr. 2013). "The Infancy of Infant Pain Research: The Experimental Origins of Infant Pain Denial". In: *The Journal of Pain* 14.4, pp. 338–350. ISSN: 15265900. DOI: 10.1016/j.jpain.2012.12.017 (cited page 15).

Rodrigues, Johannes et al. (2021). "EPOS: EEG Processing Open-Source Scripts". In: *Frontiers in Neuroscience* 15. ISSN: 1662-453X (cited pages 120–122).

Roofthooft, Daniëlla W.E. et al. (2014). "Eight Years Later, Are We Still Hurting Newborn Infants?" In: *Neonatology* 105.3, pp. 218–226. ISSN: 1661-7800, 1661-7819. DOI: 10.1159/000357207 (cited page 16).

Rosner, Bernard (Mar. 2012). "Percentage Points for a Generalized ESD Many-Outlier Procedure". In: *Technometrics* (cited page 48).

Roué, Jean-Michel et al. (Oct. 2018). "Multi-Modal Pain Assessment: Are near-Infrared Spectroscopy, Skin Conductance, Salivary Cortisol, Physiologic Parameters, and Neonatal Facial Coding System Interrelated during Venepuncture in Healthy, Term Neonates?" In: *Journal of Pain Research* 11, pp. 2257–2267. ISSN: null. DOI: 10.2147/JPR.S165810 (cited pages 31, 114, 116).

Ruiz de Miras, Juan (2016). "Fractal Analysis in MATLAB: A Tutorial for Neuroscientists". In: *The Fractal Geometry of the Brain*. Ed. by Antonio Di Ieva. Springer Series in Computational Neuroscience. New York, NY: Springer, pp. 523–532. ISBN: 978-1-4939-3995-4. DOI: 10.1007/978-1-4939-3995-4_33 (cited page 54).

Rupawala, Mohammed et al. (Apr. 2022). *Developmental Switch in Prediction and Adaptation to Pain in Human Neonates*. DOI: 10.1101/2022.04.05.486988 (cited pages 49, 112, 116).

— (Apr. 2023). "A Developmental Shift in Habituation to Pain in Human Neonates". In: *Current Biology* 33.8, 1397–1406.e5. ISSN: 09609822. DOI: 10.1016/j.cub.2023.02.071 (cited pages 35, 49, 112).

Sadek, Ibrahim et al. (2017). "Continuous and Unconstrained Vital Signs Monitoring with Ballistocardiogram Sensors in Headrest Position". In: *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Orland, FL, USA: IEEE, pp. 289–292. ISBN: 978-1-5090-4179-4. DOI: 10.1109/BHI.2017.7897262 (cited pages 90, 106).

Salekin, Md Sirajus et al. (2021a). "Future Roles of Artificial Intelligence in Early Pain Management of Newborns". In: *Paediatric and Neonatal Pain* 3.3, pp. 134–145. ISSN: 2637-3807. DOI: 10.1002/pne2.12060 (cited page 31).

Salekin, Md Sirajus et al. (Feb. 2021b). "Multimodal Spatio-Temporal Deep Learning Approach for Neonatal Postoperative Pain Assessment". In: *Computers in Biology and Medicine* 129, p. 104150. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2020.104150. arXiv: 2012.02175 (cited pages 114, 116).

Sanders, L.J. (2004). "The Charcot Foot: Historical Perspective 1827-2003". In: *Diabetes/Metabolism Research and Reviews* 20.SUPPL. 1, S4–S8. ISSN: 1520-7552. DOI: 10.1002/dmrr.451 (cited page 12).

Schenk, Karin et al. (Dec. 2019). "The Influence of Gestational Age in the Psychometric Testing of the Bernese Pain Scale for Neonates". In: *BMC Pediatr* 19.1, p. 20. ISSN: 1471-2431. DOI: 10.1186/s12887-018-1380-8 (cited page 29).

Scher, Mark S. (May 1997). "Neurophysiological Assessment of Brain Function and Maturation II. A Measure of Brain Dysmaturity in Healthy Preterm Neonates". In: *Pediatric Neurology* 16.4, pp. 287–295. ISSN: 08878994. DOI: 10.1016/S0887-8994(96)00009-4 (cited page 38).

Schmidt Mellado, Gabriela et al. (Jan. 2022). "The Impact of Premature Extrauterine Exposure on Infants' Stimulus-Evoked Brain Activity across Multiple Sensory Systems". In: *NeuroImage: Clinical* 33, p. 102914. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2021.102914 (cited pages 40, 44, 45, 49, 53, 68, 85, 116).

Sellam, Gila et al. (Aug. 2011). "Review: Contextual Factors Influencing Pain Response to Heelstick Procedures in Preterm Infants: What Do We Know? A Systematic Review". In: *European Journal of Pain* 15.7. ISSN: 1090-3801, 1532-2149. DOI: 10.1016/j.ejpain.2011.01.002 (cited page 75).

Serpa, Ana Beatriz Mello et al. (Jan. 2007). "Multidimensional Pain Assessment of Preterm Newborns at the 1st, 3rd and 7th Days of Life". In: *Sao Paulo Med. J.* 125.1, pp. 29–33. ISSN: 1516-3180. DOI: 10.1590/S1516-31802007000100006 (cited page 29).

Seymour, Ben (Mar. 2019). "Pain: A Precision Signal for Reinforcement Learning and Control". In: *Neuron* 101.6, pp. 1029–1041. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2019.01.055 (cited page 13).

Seymour, Ben and Flavia Mancini (Nov. 2020). "Hierarchical Models of Pain: Inference, Information-Seeking, and Adaptive Control." In: *NeuroImage* 222, p. 117212. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117212 (cited page 13).

Shimauchi, Suehiro et al. (2021). "R-R Interval Estimation for Wearable Electrocardiogram Based on Single Complex Wavelet Filtering and Morphology-Based Peak Selection". In: *IEEE Access* 9, pp. 60802–60827. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3070604 (cited pages 90, 106).

Siegel, Allan and Hreday N Sapru (2006). *Essential Neuroscience*. Lippincott Williams & Wilkins (cited pages 14, 31).

Simons, Laura E. et al. (2014). "The Human Amygdala and Pain: Evidence from Neuroimaging". In: *Human Brain Mapping* 35.2, pp. 527–538. ISSN: 1097-0193. DOI: 10.1002/hbm.22199 (cited page 16).

Slater, Rebeccah (Apr. 2006). "Cortical Pain Responses in Human Infants". In: *Journal of Neuroscience* 26.14, pp. 3662–3666. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.0348-06.2006 (cited pages 32, 37).

Slater, Rebeccah et al. (2010a). "Evoked Potentials Generated by Noxious Stimulation in the Human Infant Brain". In: *European Journal of Pain* 14.3, pp. 321–326. ISSN: 1532-2149. DOI: 10.1016/j.ejpain.2009.05.005 (cited pages 7, 23, 34, 53, 126).

Slater, Rebeccah et al. (2010b). "Oral Sucrose as an Analgesic Drug for Procedural Pain in Newborn Infants: A Randomised Controlled Trial". In: 376, p. 8 (cited pages 37, 50, 68).

Slater, Rebeccah et al. (Aug. 2010c). "Premature Infants Display Increased Noxious-Evoked Neuronal Activity in the Brain Compared to Healthy Age-Matched Term-Born Infants". In: *NeuroImage* 52.2, pp. 583–589. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2010.04.253 (cited page 34).

Slater, Rebeccah et al. (Sept. 2020). "Preterm Health: Time to Bridge the Evidence Gap". In: *The Lancet* 396.10255, pp. 872–873. ISSN: 01406736. DOI: 10.1016/S0140-6736(20)31977-2 (cited pages 6, 19, 30).

Slater, Rebeccah et al. (May 2023). "Systems and Methods for Measuring a Response of a Subject to an Event". WO2023084208A1 (cited page 74).

Smith, Stephen M. et al. (Jan. 2004). "Advances in Functional and Structural MR Image Analysis and Implementation as FSL". In: *NeuroImage*. Mathematics in Brain Imaging 23, S208–S219. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2004.07.051 (cited page 49).

Smith, Steven W (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. 2nd ed. California Technical Publishing. ISBN: 0-9660176-6-8 (cited pages 131, 132).

Sternberg, Wendy F. et al. (Feb. 2005). "Long-Term Effects of Neonatal Surgery on Adulthood Pain Behavior". In: *Pain* 113.3, pp. 347–353. ISSN: 0304-3959. DOI: 10.1016/j.pain.2004.11.013 (cited page 18).

Stroud, K. A. and Dexter J. Booth (Mar. 2020). *Advanced Engineering Mathematics*. Bloomsbury Publishing. ISBN: 978-1-352-01025-1 (cited page 22).

Suliman, Ahmad et al. (2019). "Performance Comparison for Ballistocardiogram Peak Detection Methods". In: *IEEE Access* 7, pp. 53945–53955. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2912650 (cited pages 90, 106).

Sun, Xu and Weichao Xu (Nov. 2014). "Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves". In: *IEEE Signal Processing Letters* 21.11, pp. 1389–1393. ISSN: 1558-2361. DOI: 10.1109/LSP.2014.2337313 (cited page 94).

Taktak, Azzam F. G. et al., eds. (2014). *Clinical Engineering: A Handbook for Clinical and Biomedical Engineers*. First edition. Amsterdam Waltham, MA: Elsevier : Academic Press. ISBN: 978-0-12-396961-3 978-0-12-397238-5 (cited page 28).

Talebi, Safa, Javad Frounchi, and Behzad Mozaffari Tazehkand (Dec. 2022). *A Novel Channel Selection Approach for Human Neonate's Pain EEG Data Analysis*. DOI: 10.21203/rs.3.rs-2390234/v1 (cited page 38).

Tivadar, Ruxandra I. and Micah M. Murray (Jan. 2019). "A Primer on Electroencephalography and Event-Related Potentials for Organizational Neuroscience". In: *Organizational Research Methods* 22.1, pp. 69–94. ISSN: 1094-4281, 1552-7425. DOI: 10.1177/1094428118804657 (cited pages 21, 53).

Tracey, Irene and Patrick W. Mantyh (Aug. 2007). "The Cerebral Signature for Pain Perception and Its Modulation". In: *Neuron* 55.3, pp. 377–391. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.07.012 (cited pages 7, 14).

Tracey, Irene et al. (Apr. 2002). "Imaging Attentional Modulation of Pain in the Periaqueductal Gray in Humans". In: *J. Neurosci.* 22.7, pp. 2748–2752. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.22-07-02748.2002 (cited page 14).

van den Bosch, Gerbrich E. et al. (2015). "Prematurity, Opioid Exposure and Neonatal Pain: Do They Affect the Developing Brain?" In: *Neonatology* 108.1, pp. 8–15. ISSN: 1661-7800, 1661-7819. DOI: 10.1159/000376566 (cited pages 6, 18, 110, 111).

van der Vaart, Marianne et al. (Sept. 2019). "Multimodal Pain Assessment Improves Discrimination between Noxious and Non-noxious Stimuli in Infants". In: *Paediatric and Neo Pain* 1.1, pp. 21–30. ISSN: 2637-3807, 2637-3807. DOI: 10.1002/pne2.12007 (cited pages 40, 44, 45, 68, 76, 82, 84, 90, 105, 109, 111, 116).

van der Vaart, Marianne et al. (Aug. 2022). "Premature Infants Display Discriminable Behavioral, Physiological, and Brain Responses to Noxious and Nonnoxious Stimuli". In: *Cereb Cortex* 32.17, pp. 3799–3815. ISSN: 1047-3211. DOI: 10.1093/cercor/bhab449 (cited page 46).

Vandenbroucke, Jan P. et al. (Nov. 2007). "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration". In: *Epidemiology* 18.6, p. 805. ISSN: 1044-3983. DOI: 10.1097/EDE.0b013e3181577511 (cited page 144).

Vanhatalo, Sampsa and Leena Lauronen (Dec. 2006). "Neonatal SEP – Back to Bedside with Basic Science". In: *Seminars in Fetal and Neonatal Medicine*. Assessing Brain Function in the Perinatal Period 11.6, pp. 464–470. ISSN: 1744-165X. DOI: 10.1016/j.siny.2006.07.009 (cited pages 76, 106).

Vederhus, Bente Johanne et al. (Oct. 2012). "Pain Tolerance and Pain Perception in Adolescents Born Extremely Preterm". In: *The Journal of Pain* 13.10, pp. 978–987. ISSN: 1526-5900, 1528-8447. DOI: 10.1016/j.jpain.2012.07.008 (cited page 18).

Verriotis, Madeleine et al. (Feb. 2015). "Cortical Activity Evoked by Inoculation Needle Prick in Infants up to One-Year Old". In: *Pain* 156.2, pp. 222–230. ISSN: 0304-3959. DOI: 10.1097/01.j.pain.0000460302.56325.0c (cited page 34).

Victoria, Nicole C. and Anne Z. Murphy (Feb. 2016). "Exposure to Early Life Pain: Long Term Consequences and Contributing Mechanisms". In: *Curr Opin Behav Sci* 7, pp. 61–68. ISSN: 2352-1554. DOI: 10.1016/j.cobeha.2015.11.015 (cited pages 17, 108, 110).

Villringer, A. et al. (May 1993). "Near Infrared Spectroscopy (NIRS): A New Tool to Study Hemodynamic Changes during Activation of Brain Function in Human Adults". In: *Neuroscience Letters* 154.1-2, pp. 101–104. ISSN: 03043940. DOI: 10.1016/0304-3940(93)90181-J (cited page 32).

Vinall, Jillian et al. (Sept. 2013). "Parent Behaviors Moderate the Relationship between Neonatal Pain and Internalizing Behaviors at 18 Months Corrected Age in Children Born Very Prematurely". In: *PAIN* 154.9, pp. 1831–1839. ISSN: 0304-3959. DOI: 10.1016/j.pain.2013.05.050 (cited page 6).

Vinall, Jillian et al. (Mar. 2014). "Invasive Procedures in Preterm Children: Brain and Cognitive Development at School Age". In: *Pediatrics* 133.3, pp. 412–421. ISSN: 0031-4005, 1098-4275. DOI: 10.1542/peds.2013-1863 (cited pages 18, 108, 110).

Von Bargen, Donna M. (1983). "Infant Heart Rate: A Review of Research and Methodology". In: *Merrill-Palmer Quarterly* 29.2, pp. 115–149. ISSN: 0272-930X. JSTOR: 23086214 (cited page 91).

Wadman, Wytse J. and Fernando H. Lopes da Silva (Nov. 2017). "C4Biophysical Aspects of EEG and MEG Generation". In: *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Ed. by Donald L. Schomer et al. Oxford University Press, p. 0. ISBN: 978-0-19-022848-4. DOI: 10.1093/med/9780190228484.003.0004 (cited page 20).

Wager, Tor D. et al. (Apr. 2013). "An fMRI-Based Neurologic Signature of Physical Pain". In: *N Engl J Med* 368.15, pp. 1388–1397. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1204471 (cited pages 17, 33).

Walco, Gary A. et al. (Feb. 2018). "Clinical Trial Designs and Models for Analgesic Medications for Acute Pain in Neonates, Infants, Toddlers, Children, and Adolescents: ACTTION Recommendations". In: *PAIN* 159.2, p. 193. ISSN: 0304-3959. DOI: 10.1097/j.pain.0000000000001104 (cited page 81).

Walton, Marc K. et al. (Sept. 2015). "Clinical Outcome Assessments: Conceptual Foundation—Report of the ISPOR Clinical Outcomes Assessment – Emerging Good Practices for Outcomes Research Task Force". In: *Value in Health* 18.6, pp. 741–752. ISSN: 10983015. DOI: 10.1016/j.jval.2015.08.006 (cited page 24).

Wayant, Cole, Jared Scott, and Matt Vassar (Nov. 2018). "Evaluation of Lowering the P Value Threshold for Statistical Significance From .05 to .005 in Previously Published Randomized Clinical Trials in Major Medical Journals". In: *JAMA* 320.17, pp. 1813–1815. ISSN: 0098-7484. DOI: 10.1001/jama.2018.12288 (cited page 104).

Wiech, Katja et al. (Aug. 2014). "Influence of Prior Information on Pain Involves Biased Perceptual Decision-Making". In: *Current Biology* 24.15, R679–R681. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.06.022 (cited page 12).

Wiedemann, H. -R. (Oct. 1994). "Hans Berger". In: *Eur J Pediatr* 153.10, pp. 705–705. ISSN: 1432-1076. DOI: 10.1007/BF01954482 (cited page 19).

Williams, Morika D. and B. Duncan X. Lascelles (Feb. 2020). "Early Neonatal Pain—A Review of Clinical and Experimental Implications on Painful Conditions Later in Life". In: *Front Pediatr* 8, p. 30. ISSN: 2296-2360. DOI: 10.3389/fped.2020.00030 (cited pages 6, 17, 108, 110).

Winkler, Irene, Stefan Haufe, and Michael Tangermann (Aug. 2011). "Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals". In: *Behavioral and Brain Functions* 7.1, p. 30. ISSN: 1744-9081. DOI: 10.1186/1744-9081-7-30 (cited pages 48, 53, 54).

Winkler, Irene et al. (May 2014). "Robust Artifactual Independent Component Classification for BCI Practitioners". In: *J. Neural Eng.* 11.3, p. 035013. ISSN: 1741-2552. DOI: 10.1088/1741-2560/11/3/035013 (cited page 53).

Wollgarten-Hadamek, Iris et al. (Apr. 2011). "Do School-Aged Children with Burn Injuries during Infancy Show Stress-Induced Activation of Pain Inhibitory Mech-

anisms?" In: *Eur J Pain* 15.4, 423.e1–10. ISSN: 1532-2149. DOI: 10.1016/j.ejpain.2010.09.001 (cited page 18).

Woo, Choong-Wan et al. (Feb. 2017). "Quantifying Cerebral Contributions to Pain beyond Nociception". In: *Nat Commun* 8.1, p. 14211. ISSN: 2041-1723. DOI: 10.1038/ncomms14211 (cited pages 17, 33).

Woody, Charles D. (Nov. 1967). "Characterization of an Adaptive Filter for the Analysis of Variable Latency Neuroelectric Signals". In: *Med. & biol. Engng.* 5.6, pp. 539–554. ISSN: 1741-0444. DOI: 10.1007/BF02474247 (cited pages 24, 85).

Woolf, Clifford J. (Nov. 2010). "What Is This Thing Called Pain?" In: *J Clin Invest* 120.11, pp. 3742–3744. ISSN: 0021-9738. DOI: 10.1172/JCI45178 (cited page 110).

Worley, A. et al. (Apr. 2012). "Multi-Modal Pain Measurements in Infants". In: *Journal of Neuroscience Methods* 205.2, pp. 252–257. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2012.01.009 (cited pages 22, 41, 43–45).

Zamzmi, Ghada et al. (Dec. 2016). "An Approach for Automated Multimodal Analysis of Infants' Pain". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. Cancun: IEEE, pp. 4148–4153. ISBN: 978-1-5090-4847-2. DOI: 10.1109/ICPR.2016.7900284 (cited pages 29, 109, 114).

Zamzmi, Ghada et al. (2017). "Automated Pain Assessment in Neonates". In: *Image Analysis*. Ed. by Puneet Sharma and Filippo Maria Bianchi. Vol. 10270. Cham: Springer International Publishing, pp. 350–361. ISBN: 978-3-319-59128-5 978-3-319-59129-2. DOI: 10.1007/978-3-319-59129-2_30 (cited page 109).

Zamzmi, Ghada et al. (2018). "A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases". In: *IEEE Reviews in Biomedical Engineering* 11, pp. 77–96. ISSN: 1941-1189. DOI: 10.1109/RBME.2017.2777907 (cited page 29).