THE UNIVERSITY *of* EDINBURGH

Longitudinal assessment of the impact of preterm birth on cognitive functions and identification of novel methods for stratification of children at risk of long-term impairment

By

Evdoxia Valavani

Thesis submitted in fulfilment of the requirements for a

Master By Research Degree

Supervisors: Prof A. Tsanas, Prof J. Boardman, Prof R. Chin, Dr. D. MacIntyre

The University of Edinburgh

2024

ABSTRACT

During the last few decades advances in perinatal care have led to significant improvements in the survival of preterm infants. However, survivors of preterm birth are at increased risk of brain maldevelopment, and subsequent neurodevelopmental deficits compared to their term-born peers. Given that neurodevelopmental trajectories are amenable to early interventions, there is a critical window of opportunity to improve long-term outcomes for children born preterm. Currently, there are no clinical tools which accurately predict brain growth or subsequent neurodevelopmental outcomes for preterm infants. Hence, the objectives of this study were: (a) to identify the early life exposures that impact total and regional brain volumes at term-equivalent age following preterm birth, and (b) to accurately predict language outcomes at two years Corrected Gestational Age (CGA). To this end, we analysed data from a longitudinal cohort of preterm infants born before 33 weeks of gestation at the Royal Infirmary of Edinburgh and developed parsimonious machine learning models using feature selection and advanced machine learning techniques. This work revealed that a combination of clinical, biological, and environmental exposures, including potentially modifiable risk factors, such as postnatal nutrition, respiratory illness, and socioeconomic deprivation, best predicts cerebral tissue volumes at term-equivalent age. Furthermore, we found that a combination of diffusion tensor imaging features and clinical perinatal factors collected as part of routine care accurately predict language outcomes at two years CGA. These results have important implications for clinical practice: mitigating these risk factors can inform current perinatal practices, leading to enhanced brain development and neurodevelopmental outcomes following preterm birth. Moreover, these findings could facilitate timely identification of infants

who are at considerable risk of language impairment and who may benefit from targeted early interventions and support services. Overall, our research can potentially offer preterm infants a healthier start in life, improved long-term outcomes and a better quality of life.

LAY SUMMARY

In recent years, advances in medical care have significantly increased the survival rates of preterm infants (i.e., babies who are born before completing 37 weeks of pregnancy). However, these infants still face a higher risk of impaired brain growth and subsequent neurodevelopmental problems compared to term babies (born after 37 weeks of pregnancy). Although interventions during early childhood can significantly improve outcomes for preterm infants, there are currently no clinical tools to reliably predict brain growth or neurodevelopmental outcomes for these babies. This research aimed to: (a) identify the early life exposures that affect brain growth, and thus, the brain size of preterm infants, and (b) develop a tool that predicts language abilities at two years of age. To achieve these goals, we studied a group of preterm infants who were born before 33 weeks of pregnancy at the Royal Infirmary of Edinburgh. We found that a set of clinical, biological, and environmental exposures predicts brain volumes in preterm infants. Some of these exposures, such as the infant's nutrition, respiratory illness, and socioeconomic deprivation, can potentially be modified to enhance brain growth. In addition, we found that another set of clinical and imaging features during early life can reliably predict language abilities at two years of age. These findings are important because they can inform current clinical practices, leading to enhanced brain development and better outcomes for preterm infants. In addition, this research can help clinicians to timely detect which preterm infants are at risk of future language problems, allowing for targeted interventions and support services. Overall, our research can offer preterm infants a healthier life start, improved long-term outcomes, and ultimately, a better quality of life.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my supervisors Prof Athanasios Tsanas, Prof James P Boardman, Prof Richard Chin, and Dr Donald MacIntyre for their guidance, invaluable advice, continuous support and incredible patience throughout my studies. Special thanks go to my primary supervisor, prof Athanasios Tsanas, for believing in me from the start and for the funding opportunity that allowed me to conduct this thesis. My sincere thanks also go to the members of my thesis committee, Prof Steff Lewis, Dr Kathrin Cresswell, and Dr Sinead Rhodes for their insightful comments and suggestions, and helpful feedback. I would like to extend my sincere gratitude to Dr Patrick Hadoke for his motivation, support, and encouragement. Additionally, I would like to thank all researchers and colleagues who are part of the Theirworld Edinburgh Birth Cohort, as well as all participating families, without whom this study would not have been possible.

I would also like to thank my mum and my sister for always believing in me, for their unwavering love, continuous encouragement and constant support during my studies. Special thanks also go to my friends for supporting me especially during the COVID-19 pandemic. Last but not least, I would like to thank my partner, George, for his love, patience, and understanding, and for constantly motivating me to write this thesis.

ABBREVIATIONS

Bayley-III	Bayley Scales of Infant and Toddler Development, Third Edition
BMI	Body Mass Index
BPD	Bronchopulmonary Dysplasia
CGA	Corrected Gestational Age
CI	Confidence Interval
dMRI	Diffusion Magnetic Resonance Imaging
DTI	Diffusion Tensor Imaging
GA	Gestational Age
IQ	Intelligence Quotient
IQR	Interquartile Range
LASSO	Least Absolute Shrinkage and Selection Operator
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
$MgSO_4$	Magnesium Sulphate
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NEC	Necrotizing Enterocolitis
NICE	National Institute for Health and Care Excellence
NICU	Neonatal Intensive Care Unit

PDP	Partial Dependence Plot
RF	Random Forests
RMSE	Root Mean Squared Error
ROP	Retinopathy Of Prematurity
SIMD2016	Scottish Index of Multiple Deprivation 2016
sMRI	Structural Magnetic Resonance Imaging
TEBC	Theirworld Edinburgh Birth Cohort
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Background

Preterm birth is defined by the World Health Organization (WHO) as any live birth that occurs before 37 completed weeks of gestation and can be further subcategorised based on Gestational Age (GA) into four groups: late preterm (GA between 34 weeks and 36 weeks and 6 days), moderate preterm (GA between 32 weeks and 33 weeks and 6 days), very preterm (GA between 28 weeks and 31 weeks and 6 days), and extremely preterm (GA less than 28 weeks) (WHO, 2018). Worldwide, an estimated 13.4 million infants are born preterm annually, indicating a global prevalence of about 10% (Ohuma et al., 2023). Although in resource-rich settings, advances in perinatal care have led to marked improvements in the survival of preterm infants over recent decades (Mehler et al., 2016; Myrhaug et al., 2019; Norman et al., 2019; Patel et al., 2017; Stoll et al., 2015), preterm birth remains the leading cause of infant and childhood mortality. In 2020, preterm birth and its associated complications accounted for 17.7% of all deaths in children under 5 years old, and 36% of neonatal deaths, globally (Perin et al., 2022).

The preterm brain is particularly vulnerable to insults that occur prenatally and during the early postnatal period, which can lead to altered brain development. Survivors of preterm birth are at increased risk of life-long neurodevelopmental adverse outcomes compared to children born at full term. Moreover, there is an inverse relationship between the risk of morbidity and GA, with extremely preterm infants being at the

highest risk, followed by very preterm infants, and then moderate and late preterm infants (Fernández de Gamarra-Oca et al., 2021; Pascal et al., 2018; Sarda et al., 2021). Adverse neurodevelopmental outcomes associated with prematurity include global and specific learning difficulties, executive dysfunction, inattentiveness, motor impairments, deficits in language skills, hearing and visual disorders, and poor social, emotional, and behavioural functioning. Many mental health disorders are also more frequent among adolescents and adults who were born preterm (S. Johnson & Marlow, 2017).

Most information about the impact of preterm birth on brain development comes from older studies that do not reflect recent advances in perinatal care (Boardman et al., 2020). Thus, there is a need to analyse data that comes from a 21st-century cohort in order to investigate the early life risk factors that affect a child's growing brain and subsequent neurodevelopmental outcomes. The underlying rationale is that these insights may lead to the development of new methods for the timely detection of children who are at greatest risk for long–term neurodevelopmental impairment and who may benefit from targeted early developmental interventions. In addition, identifying the key factors that affect brain growth following preterm birth could further improve current perinatal treatments and neuroprotective strategies, which could potentially lead to better long–term outcomes and quality of life for this vulnerable population.

1.2. Statement of the problem

Increased survival rates following preterm birth are accompanied by high incidence rates of a wide range of negative sequelae. Approximately one-third of children born very preterm will have neurodevelopmental impairment at 2 years of age (Trittmann, Nelin and Klebanoff, 2013; Kidokoro et al., 2014). Around 5% to 10% of children born preterm will have cerebral palsy, 25% to 40% will suffer from other motor impairments, and 25% to 50% will have cognitive deficits, attentional, social, emotional, and behavioural problems, which persist into adulthood (Crump, 2020; Volpe, 2019), substantially diminishing their quality of life and imposing a significant economic burden to affected individuals, their families and caregivers, as well as high health and educational costs to society (Hua et al., 2023; Mangham et al., 2009).

Little is known about the biological, environmental, and social factors that affect brain development and subsequent neurodevelopmental outcomes following preterm birth, which presents challenges for risk stratification and for the discovery and evaluation of neuroprotective strategies. Neurodevelopmental trajectories are amenable to early developmental interventions, which presents a window of opportunity to have a profound, long-lasting effect on later life (Orton et al., 2024). However, currently, there is a lack of clinical tools which accurately predict long–term neurodevelopmental outcomes for children born preterm. Thus, there is a clear unmet clinical need for the identification of the early life risk factors that impact the brain development of the preterm infant, and early detection of those children who are at high risk of adverse outcomes. Timely identification will assist in targeting intervention programmes which can potentially prevent or mitigate later challenges, thus improving development, and quality of life.

1.3. Purpose of the study

The overall objective of the current study was the identification of the early life exposures that impact brain development and subsequent neurodevelopment, and the development of novel and multifactorial tools for stratification of children at risk of long–term neurodevelopmental impairments following preterm birth. Identifying potentially modifiable risk factors that affect brain development, and timely detection of children at high risk of long–term deficits, will give health professionals the best opportunity to treat and intervene early to improve outcomes for this vulnerable population. The ultimate goal is to offer preterm infants a healthier life start, improved long–term functioning and a better quality of life.

1.4. Research objectives and hypotheses

The first objective of the current thesis was to identify the predictors of altered brain growth associated with preterm birth. Specifically, we aimed to explore the early life risk factors that jointly impact preterm brain growth and develop a machine learning model to accurately predict global and regional cerebral tissue volumes at term–equivalent age following preterm birth. Our goal was to identify the combination of the most important perinatal variables that are associated with the neuroanatomical abnormalities underlying neurodevelopmental deficits following preterm birth and to identify potentially modifiable risk factors to enhance perinatal treatments and interventions, ultimately improving long–term neurodevelopmental outcomes and quality of life for children born preterm. We hypothesised that a combination of biological, clinical, and environmental perinatal variables would accurately predict

global and regional cerebral volumes following preterm birth. Given that different brain regions are characterised by different growth rates, we also hypothesised that they are differentially vulnerable to perinatal exposures.

The second objective of this thesis was to develop a machine learning model that accurately predicts language outcomes at two years of CGA following preterm birth. We hypothesised that a model which combines clinical, environmental, and brain imaging features that capture generalised white matter dysmaturation, would outperform existing statistical models, thus allowing for accurate prediction of language outcomes at two years CGA and timely identification of preterm infants who are at greatest risk of language deficits and who would benefit most from targeted early interventions and support services.

1.5. Structure of the thesis

This thesis is organised into seven chapters. Chapter 1 introduces the research topic, provides background information, outlines the overarching goals of the study, and presents an overview of the thesis structure. Chapter 2 provides a review of the relevant literature, identifies knowledge gaps, and outlines the specific research objectives and hypotheses addressed by this study. Chapter 3 describes the study design and data collection methods. Chapter 4 provides a detailed overview of the methodology used for data analysis. Chapter 5 presents the findings on prediction of global and regional cerebral volumes following preterm birth. Chapter 6 presents a machine learning model for prediction of language outcome at two years corrected gestational age (CGA) following preterm birth. Finally, Chapter 7 summarises and

provides some overall context for the main findings of the study, outlines the limitations, and outlines areas for future research.

CHAPTER 2: LITERATURE REVIEW

2.1. The preterm brain

2.1.1. Preterm brain development and neuroimaging findings

The third trimester of gestation is characterised by accelerated foetal brain development in terms of both structure and function (Kapellou et al., 2006). During this critical period, several biological processes take place, including a rapid increase in brain volume, the formation of sulci and gyri, neuronal differentiation and organisation, myelination, spinogenesis, and synaptogenesis (Kostović & Jovanov-Milošević, 2006; Kostovic & Vasung, 2009; Rajagopalan et al., 2011; Volpe, 2018). It is well established that during the late foetal human development, there is an exponential increase in total and regional cerebral tissue volumes, with different brain regions exhibiting different growth rates. Specifically, between the 25^{th} and 36^{th} weeks of pregnancy, the cerebellum exhibits the highest growth rate with a four-fold increase in volume, while the total brain and regional cerebral volumes are characterised by a two-fold increase in size (Andescavage et al., 2016; Clouchoux et al., 2012).

Preterm birth results in an abrupt and premature exposure of the developing foetal brain to the extrauterine environment during a period of critical biological processes (Kapellou et al., 2006), which renders the brain particularly vulnerable to insults associated with preterm birth, such as ischaemia, inflammation, excitotoxicity, and oxidative stress, potentially leading to brain injury and dysmaturation (Volpe, 2019). Preterm birth is strongly associated with aberrant brain development, impaired

function and subsequent neurocognitive impairment (see sections 2.1.2 and 2.1.3). Neuroimaging at term-equivalent age reveals significant differences in brain development between preterm infants and their full-term counterparts. Magnetic Resonance Imaging (MRI) has been extensively used to assess preterm brain development in terms of both macrostructure and microstructure.

Structural MRI (sMRI) with conventional T1 and T2 weighted sequences provides a detailed assessment of neonatal brain anatomy (Counsell et al., 2019), and is a commonly used tool for brain morphometry and volumetric analysis of the preterm brain (Counsell et al., 2019). Groupwise studies using volumetric analysis have shown that very preterm infants tend to have smaller total and regional brain volumes at term-equivalent age compared to full-term infants. Specifically, the preterm brain is characterised by reduced volumes of white matter, cortical grey matter, deep grey matter structures, and the cerebellum (Ball et al., 2012; Batalle et al., 2018; Boardman et al., 2006; Inder et al., 2005; Limperopoulos et al., 2005; Loh et al., 2017; Mewes et al., 2006; Srinivasan et al., 2007; Thompson et al., 2007). However, growth failure is not inevitable (Boardman et al., 2007) suggesting that considerable individual variation exists.

In addition to the anatomic details provided by sMRI, information about the microstructural organisation and connectivity of the developing brain can be obtained by Diffusion MRI (dMRI), specifically Diffusion Tensor Imaging (DTI). This type of MRI measures the diffusion of water molecules in brain tissues over time, providing information about the brain's microstructure (Counsell et al., 2019). In the cerebrospinal fluid, water molecules are not restricted and are free to move in any

direction, meaning that the diffusion is *isotropic* (i.e., water molecules diffuse equally in all directions). Conversely, in the white matter which mainly consists of axons, water movement/diffusion is restricted along the direction of the axons, making the diffusion anisotropic (Counsell et al., 2019). Typical DTI measures used to describe the microstructural integrity of cerebral white matter, including water content, degree of myelination, neuronal density, and axon integrity and orientation, involve Fractional Anisotropy (FA – refers to the degree of anisotropy of water diffusion within the brain tissue), Axial Diffusivity (AD – measures water diffusion parallel to axon tracts), Radial Diffusivity (RD – measures water diffusion perpendicular to axon tracts), and Mean Diffusivity (MD – measure the average diffusion of water molecules) (Counsell et al., 2019). Multiple studies have shown that at term-equivalent age, preterm infants exhibit altered white matter microstructure compared to infants born at term. Specifically, the white matter tracts of the preterm brain usually exhibit lower values of FA, and higher values of MD, AD, and RD compared to full-term neonates, indicative of diffuse white matter tissue damage (Batalle et al., 2018; Brossard-Racine et al., 2017; Dibble et al., 2021; Kaur et al., 2014; Kelly et al., 2019; Knight et al., 2018; Pannek et al., 2018; Pecheva et al., 2018; Pogribna et al., 2013; S. E. Rose et al., 2008; Thompson et al., 2011; Thompson et al., 2019a).

Although the DTI parameters mentioned above can reliably capture microstructural white matter alterations in the preterm brain, whole-brain calculation of these metrics is computationally expensive, making it difficult to use them in models as early biomarkers of future neurocognitive impairment. Thus, Blesa et al. demonstrated that another set of metrics derived from histogram analysis of DTI parameters of the skeletonised white matter tracts (i.e., peak width of skeletonised [PS] -FA, -MD, -RD,

and -AD) can be used to assess white matter microstructure in preterm infants (Blesa et al., 2020). At term-equivalent age, preterm infants tend to have higher values of peak width of skeletonised DTI metrics compared to their term-born peers, indicative of white matter dysmaturation. The advantage of the histogram-based framework is that it is computationally inexpensive and has high inter-scanner reproducibility, making it suitable for clinical settings and multi-centre studies (Blesa et al., 2020).

2.1.2. Impact of perinatal factors on preterm brain development

Several perinatal risk factors have been associated with atypical brain development and subsequent neurodevelopmental deficits in preterm infants. First of all, multiple studies have demonstrated that an increasing degree of prematurity is associated with smaller global and regional cerebral tissue volumes as well as decreased white matter maturation characterised by lower FA and higher MD values of white matter tracts on DTI (Anjari et al., 2009; Ball et al., 2010, 2012; Boardman et al., 2006; Inder et al., 2005; Kidokoro et al., 2014; Limperopoulos et al., 2005; Partridge et al., 2004). Birthweight is strongly associated with preterm brain volume and microstructural integrity, with increasing values associated with greater global and regional cerebral volumes and more mature white matter (Alexander et al., 2019; Knickmeyer et al., 2016; Matthews et al., 2018; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Thompson et al., 2019a). Intrauterine growth restriction has also been associated with reduced total and regional brain volumes, and alterations in white matter microstructure and brain connectivity (Boardman & Counsell, 2020; S. L. Miller et al., 2016). Moreover, sex also impacts white matter microstructure, as well as brain volumes, with males having larger global brain volumes than females (Alexander et

al., 2019; Dibble et al., 2021; Gilmore et al., 2007; Kersbergen et al., 2016; Matthews et al., 2018; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Ruigrok et al., 2014; Thompson et al., 2007; Thompson et al., 2019a).

Additional maternal and foetal factors that negatively affect the growth and white matter microstructural integrity of the developing brain include lack of antenatal steroids (Pogribna et al., 2013; Rogers et al., 2016), multiple birth (Alexander et al., 2019; Thompson et al., 2019b), chorioamnionitis (Anblagan et al., 2016; Boardman & Counsell, 2020; Jain et al., 2022; Pogribna et al., 2013), prenatal alcohol and drug exposure (Boardman & Counsell, 2020; Donald et al., 2024; Taylor et al., 2015), maternal smoking during pregnancy (Boardman & Counsell, 2020; Ekblad et al., 2015), maternal anxiety (Boardman & Counsell, 2020; Dean et al., 2018; Lautarescu et al., 2020), and socioeconomic deprivation (Benavente-Fernández et al., 2019; Betancourt et al., 2016; Boardman & Counsell, 2020; Ene et al., 2019; Jha et al., 2019; Knickmeyer et al., 2016; Lu et al., 2021; Thompson et al., 2019b; Triplett et al., 2022).

In addition, several factors during the early postnatal period, including co-morbidities of prematurity and perinatal practices significantly affect preterm brain development. Factors that negatively impact brain growth and (Kidokoro et al., 2014; Thompson et al., 2007) white matter microstructure include white matter injury (Kidokoro et al., 2014; Pogribna et al., 2013; Thompson et al., 2007, 2008), bronchopulmonary dysplasia (BPD) (Anjari et al., 2009; Ball et al., 2010; Boardman & Counsell, 2020; Inder et al., 2005; Kidokoro et al., 2014; Thompson et al., 2007), Retinopathy Of Prematurity (ROP) (Glass et al., 2017; Sveinsdóttir et al., 2018), necrotizing enterocolitis (NEC) (Kidokoro et al., 2014; Matthews et al., 2018; Pogribna et al., 2013; Shah et al., 2008),

sepsis (Matthews et al., 2018), and treated patent ductus arteriosus (Kidokoro et al., 2014; Rogers et al., 2016; Thompson et al., 2008). Perinatal practices that have been associated with reduced brain volumes and less mature white matter microstructure include prolonged mechanical ventilation (Boardman et al., 2007; Brouwer, Kersbergen, Van Kooij, et al., 2017; Guillot et al., 2020; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Rogers et al., 2016) and parenteral nutrition (Brouwer, Kersbergen, Van Kooij, et al., 2017; Kidokoro et al., 2014), and exposure to postnatal corticosteroids (Kidokoro et al., 2014; Thompson et al., 2008). Finally, breast milk feeding during hospitalisation in the neonatal intensive care unit (NICU) is associated with improved white matter maturation and greater cerebral volumes following preterm birth (Belfort et al., 2016; Belfort & Inder, 2022; Blesa et al., 2019; Ottolini et al., 2020; Pogribna et al., 2013; J. Schneider et al., 2018; Sullivan et al., 2022).

2.1.3. Altered brain development and neurodevelopmental outcomes following preterm birth

The macrostructural and microstructural brain abnormalities identified using sMRI and dMRI, respectively, underlie the adverse neurodevelopmental outcomes observed following preterm birth. Multiple studies have demonstrated that altered global and regional cerebral volumes and changes in white matter microstructural integrity in preterm infants are associated with a range of short- and long-term neurodevelopmental deficits. Specifically, aberrant brain development following preterm birth has been associated with poor cognitive outcomes (Brouwer, Kersbergen, Van Kooij, et al., 2017; Counsell et al., 2008; Lind et al., 2011; Peterson et al., 2003; J. Rose et al., 2015; Young et al., 2015), intellectual disability and poor

academic performance (Loh et al., 2017), language deficits (Feldman et al., 2012; Lind et al., 2010), executive dysfunction (Lind et al., 2010; Woodward et al., 2005), adverse motor outcomes and cerebral palsy (Brouwer, Kersbergen, Van Kooij, et al., 2017; Counsell et al., 2008; Inder et al., 2005; Kim et al., 2016; Lind et al., 2010; Peterson et al., 2003; Rogers et al., 2016; J. Rose et al., 2015; Setänen et al., 2016), social-emotional deficits (Rogers et al., 2016), as well as behaviour and attention problems (Murray et al., 2016).

2.2. Language

2.2.1. Language outcomes following preterm birth

Atypical brain structure and function following preterm birth are associated with a range of adverse neurodevelopmental outcomes, including cognitive and language delays (Brouwer, Kersbergen, Van Kooij, et al., 2017; Counsell et al., 2008; Feldman et al., 2012; Kojima et al., 2024; Lind et al., 2010, 2011; Peterson et al., 2003; J. Rose et al., 2015; Young et al., 2015). Children born preterm are at a greater risk of language impairment compared to their term-born counterparts, both in early childhood and during their school years and adolescence (van Noort-van der Spek et al., 2012). Typically, language deficits are part of a broader cognitive delay, as language development depends on key cognitive functions, such as processing speed, memory, and attention (Hoff Esbjørn et al., 2006; Ortiz-Mantilla et al., 2008; Wolke et al., 2008). In turn, language problems can lead to long-term consequences, including poor academic performance, mental health issues, socio-emotional and behavioural

difficulties, and unemployment (Conti-Ramsden et al., 2013; Law et al., 2009), ultimately resulting in a lower quality of life.

Multiple studies have shown that early childhood interventions – at the time of highest neural plasticity – can significantly improve language outcomes for children born preterm (Bailey et al., 2005; Khurana et al., 2020; Markkula et al., 2024; McManus et al., 2012; Orton et al., 2024; Vandormael et al., 2019). Strategies that have proven effective, have focused either on global cognitive development or specifically on language skills; typical examples include sensory stimulation through activities such as music and book reading, family-centred practices, interventions focusing on the parent-infant relationship, and social support services. This means that there is a critical window of opportunity to improve long-term outcomes for this vulnerable population.

It is important to note, however, that not all children who are born preterm will develop language deficits. Several risk and resilience factors during the perinatal period significantly affect cognitive and language development following preterm birth (see section 2.1.2). Thus, it is crucial to identify which factors are predictive of future language development in order to timely detect which infants are at high risk of poor language development and who may benefit from targeted early interventions and support services. Nevertheless, to date, only few studies have focused on prediction of language development and thus, currently, there is a lack of clinical tools which accurately predict language outcomes following preterm birth (see section 2.2.2).

2.2.2. Current prediction tools of neurodevelopment following preterm birth

Although multiple studies have investigated the associations between perinatal factors and neuroimaging findings with subsequent neurodevelopment, only few studies have developed tools for early prediction (i.e., during the perinatal period) of neurodevelopmental outcomes following preterm birth. Notably, to date, the majority of studies have not focused on specific developmental delays, but rather on prediction of global neurodevelopmental impairment, which is defined as a score below two standard deviations from the mean on standardised developmental tests, encompassing cognitive, language, and motor delays, as well as vision and hearing problems (BAPM, 2008).

Tyson et al. (2008) investigated the predictive power of a range of clinical and demographic features of a cohort of 4,446 extremely preterm infants. They found that neurodevelopmental outcome at 18 to 22 months CGA was predicted by the combination of five perinatal features, including GA, birthweight, sex, exposure or non-exposure to prenatal corticosteroids, and singleton or multiple birth (Area under the ROC Curve [AUC] of the developed model was 0.751, 95% Confidence Interval [CI] 0.735 – 0.767). Ambalavanan et al. (2012) analysed data from a cohort of 13,085 preterm infants and developed a multivariate regression model for prediction of neurodevelopmental impairment at 18 to 22 months CGA following preterm birth. Their model comprised sex, respiratory disease, and findings from cranial ultrasound, including enlarged ventricular size, periventricular leukomalacia, or porencephalic cysts (AUC = 0.74 – 0.80). Later, Vesoulis et al. (2018) developed a web-based tool for prediction of neurodevelopmental impairment at two years CGA (AUC = 0.85), by

analysing data from 154 preterm neonates born before 30 weeks of gestation. This tool comprised the following features: duration of ventilation, mode of delivery, exposure or non-exposure to antenatal corticosteroids, ROP requiring surgery, and sMRI findings at term-equivalent age, including cerebellar haemorrhage size, cerebellar haemorrhage laterality, intraventricular haemorrhage grade, and white matter injury.

Nevertheless, deficits in different developmental domains require different targeted interventions and support services. Thus, tools for accurate prediction of impairment at specific developmental domains are valuable. Recently, Demirci et al. (2024) analysed data from 1,109 very preterm infants and used machine learning techniques to predict mental (cognitive/verbal) and motor impairment at two years CGA. They found that the most important perinatal predictors were birth year, GA, birth weight, intrauterine growth restriction, exposure to antenatal Magnesium Sulphate ($MgSO_4$), duration of hospitalisation in the NICU, duration of intubation, atypical findings on cranial ultrasound, maternal age, and maternal education level. However, their models comprising only perinatal variables achieved a low balanced accuracy* of 62% and 61% for mental and motor impairment, respectively. Combining perinatal and longitudinal data derived from 19-month follow-up neurodevelopmental assessments increased the performance of the models up to 72% and 73% for mental and motor impairment, respectively. However, this means that a more accurate prediction was only possible at the age of 19 months.

* Refer to section 4.9.2 for definition of balanced accuracy.

Vassar et al. (2020) focused on language outcomes following preterm birth. They analysed near-term sMRI and DTI data from 92 very preterm neonates and developed a model for prediction of language delay at 18 to 22 months CGA. This model comprised DTI variables in three brain regions (right sagitta stratum MD, right lingual gyrus AD, and right inferior occipital gyrus MD) and achieved 89% sensitivity and 86% specificity. Finally, Ball et al. (2017) demonstrated that distinct alterations of cerebral macrostructure and microstructure in preterm infants are associated with specific clinical and environmental exposures, and these brain alterations are, in turn, associated with subsequent neurodevelopment. Specifically, language outcome at two years CGA was associated with a distinct neuroanatomic alteration, which was affected by a range of perinatal variables, including age at MRI scan, GA, birthweight, need for continuous positive airway pressure, mechanical ventilation, surfactant administration, and parenteral nutrition.

2.3. Knowledge gaps

As outlined in section 2.1.2, multiple studies have investigated the risk factors associated with total and regional brain volumes at term–equivalent age following preterm birth. While these studies have primarily used parametric methods for data analysis, the true relationship between the predictors and the response is often non-linear. Consequently, parametric models may not provide an accurate representation of the underlying relationship. To date, no studies have applied non–parametric techniques to identify the combination of early life risk factors that accurately predict cerebral tissue volumes at term–equivalent age following preterm birth.

Furthermore, there are only a few prediction tools for neurodevelopmental impairment following preterm birth in the existing literature, with most focusing on composite neurodevelopmental outcomes. However, deficits in different developmental domains require different therapies and targeted interventions. In addition, no studies have combined data from different modalities to develop models for accurate prediction of neurodevelopmental outcomes following preterm birth. Thus, there is an urgent need to combine clinical, environmental, and imaging data to improve the prediction of adverse outcomes in specific developmental domains following preterm birth, including language outcomes.

CHAPTER 3: DATA ACQUISITION

3.1. Study design

This is a single–centre prospective longitudinal cohort study (Boardman et al., 2020).

3.2. Study setting

This study was conducted at the University of Edinburgh and the Simpson Centre for Reproductive Health, which is located at the Royal Infirmary of Edinburgh, NHS Lothian, United Kingdom (Boardman et al., 2020).

3.3. Study participants

We analysed data from two separate longitudinal cohorts: the Theirworld Edinburgh Birth Cohort (TEBC) and a pilot cohort. The TEBC (also referred to as the ‘phase II’ cohort) includes 300 preterm infants born before 33 completed weeks of gestation (based on first–trimester ultrasound) and 100 healthy term controls born after 37 completed weeks of gestation at the Simpson Centre for Reproductive Health (Boardman et al., 2020). The pilot cohort (also referred to as the ‘phase I’ cohort) consists of 150 preterm neonates (≤ 33 weeks GA) and 40 term controls (≥ 37 weeks GA) born at the same health centre (Boardman et al., 2020).

For the purposes of this work, we focused exclusively on the preterm participants of the phase I and phase II cohorts. Specifically, for our first objective of predicting global

and regional cerebral volumes following preterm birth, we used data from the phase II cohort (see Chapter 5), while for our second objective regarding the prediction of language outcomes at two years CGA, we analysed data from the phase I cohort (see Chapter 6).

The exclusion criteria for the study comprised congenital anomalies, chromosomal abnormalities, congenital infections, major overt parenchymal lesions (cystic periventricular leukomalacia, haemorrhagic parenchymal infarction), and post–haemorrhagic ventricular dilatation. Infants with contraindications to MRI at 3 Tesla, as well as those with excessive movement during MRI scans, were also excluded (Boardman et al., 2020).

3.4. Clinical and sociodemographic data collection

Data regarding pregnancy, birth, neonatal care, as well as sociodemographic characteristics were derived from the mothers' and infants' electronic medical records. Additional information not routinely recorded was obtained through structured maternal interviews (Boardman et al., 2020). Following delivery, placental histopathology was performed for all preterm infants, and data on pathological features of the placenta (e.g., chorioamnionitis) were recorded (Boardman et al., 2020). The selection of clinical and sociodemographic features that were included in our analyses was guided by existing literature linking biological and environmental exposures with brain development and neurodevelopmental outcomes following preterm birth (for more details, see Chapters 5 and 6). Table 3.1. presents the definitions of the features we have used in our subsequent analyses.

Table 3.1. Definitions of clinical and sociodemographic features

Gestational age at birth	Gestational age based on first trimester ultrasound
Gestational age at MRI scan	Gestational age at near-term MRI
Birth weight	Weight of infant at birth
Birth weight z-score	A measure of birth weight standardised for age and sex – standard deviation scores for birth weight
Sex	Sex of infant participant
Intrauterine growth restriction	Fetus's weight below the 10th percentile for its gestational age
Bronchopulmonary dysplasia	Oxygen requirement at ≥36 weeks corrected gestational age
Necrotizing enterocolitis	Stages II or III based on the modified Bell's staging
Early-onset bacterial sepsis	blood stream infection occurring within 72 hours of birth with (a) bacterial pathogen isolated from blood culture, or (b) blood culture growing coagulase negative staphylococcus, along with one or more signs of generalised infection, and treatment with intravenous antibiotics for 5 or more days
Late-onset bacterial sepsis	blood stream infection occurring ≥72 hours postnatally with (a) bacterial pathogen isolated from blood culture, or (b) blood culture growing coagulase negative staphylococcus, along with one or more signs of generalised infection, and treatment with intravenous antibiotics for 5 or more days

Retinopathy of prematurity	Retinopathy of prematurity requiring laser therapy or anti-vascular endothelial growth factor treatment
Duration of intubation	Days requiring intubation whilst in the neonatal unit
Duration of parenteral nutrition	Days requiring parenteral nutrition whilst in the neonatal unit
Breast milk exposure	Number of days receiving maternal breast milk and/or donor breast milk
Chorioamnionitis	Evidence of inflammation of the membranes and chorion of the placenta based on placental histopathology
Multiple birth	Multiple pregnancy, e.g. twins, triplets
Parity	Number of pregnancies
Mode of delivery	Type of birth
Maternal age	Age of maternal participant
Maternal Body Mass Index	Mother's Body Mass Index at booking
Maternal smoking	Smoking status at booking
Maternal depression	Medical history of maternal depression
Maternal anxiety	Medical history of maternal anxiety
Antenatal corticosteroids	Mother received steroids in antenatal period; any or a complete course define as two doses 24 hours apart
Antenatal Magnesium sulphate	Magnesium sulphate given to mother in the antenatal period
Maternal education level	Mother's education level
Scottish Index of Multiple Deprivation 2016 (SIMD2016)	SIMD2016 quintile based on the postcode of the participant's current address

3.5. Bayley Scales of Infant and Toddler Development, Third edition

In this study, preterm infants underwent neurodevelopmental assessment at two years CGA as part of standard NHS follow–up of children born prematurely. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) recommends that all preterm children receive a comprehensive outcome evaluation between two to two and a half years CGA (National Guideline Alliance (UK), 2017). Significant developmental deficits can be identified reasonably well by the age of 18 to 24 months and are predictive of long–term impairments (BAPM, 2008; Breeman et al., 2015; Linsell et al., 2018). Hence, a developmental assessment at two years CGA is crucial for the detection of children who may benefit from early interventions, potentially improving future outcomes and quality of life for this vulnerable population. We need to note that it is recommended to use corrected age until the age of two years when assessing neurocognitive development in children born preterm (MacDonald & Seshia, 2015).

In preschool children (i.e., those younger than three years old), intelligence cannot be measured directly. Hence, standardised instruments are used to assess developmental functioning rather than cognitive abilities in early childhood (Albers & Grieve, 2007; Piñon, 2010). In this study, the neurodevelopmental outcomes of preterm infants were assessed with the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley–III) (Boardman et al., 2020), which is the most frequently used tool in both clinical and research settings for evaluating the general developmental level of children aged between one and 42 months (Albers & Grieve, 2007).

The original Bayley Scales were developed in 1969 by psychologist Dr. Nancy Bayley and since then, they have undergone multiple revisions, with the third edition published in 2006. The Bayley–III consists of a series of developmental play tasks and takes around 45 to 60 minutes to administer (Albers & Grieve, 2007). Its goal is to identify delays in specific developmental domains and to guide appropriate interventions and support services. The Bayley–III assesses five key developmental domains through five distinct scales: (1) Cognitive, (2) Language, (3) Motor, (4) Social–Emotional, and (5) Adaptive Behaviour.

The Cognitive Scale consists of 91 items that assess skills such as finding hidden objects, comparing masses, looking for a fallen object, and symbolic and pretend play. Several items in the Cognitive Scale aim to evaluate processing speed and problem–solving abilities, which are essential skills of cognitive competence in children (Armstrong & Agazzi, 2010).

The Language Scale comprises two subtests: the Receptive Communication and the Expressive Communication Scales. The Receptive Communication subtest consists of 49 items that assess the child's ability to understand and respond to verbal stimuli. The Expressive Communication subtest contains 48 items that evaluate a child's ability to communicate with others using verbal and non–verbal communication, such as gestures and facial expressions (Albers & Grieve, 2007; Crais, 2010).

The Motor Scale comprises the Fine Motor and the Gross Motor subtests. The Fine Motor subtest, with 66 items, assesses fine motor skills, such as eye movements, early hand and finger movements, grasping patterns, bimanual coordination, prewriting

skills, controlled finger and hand movement, and stacking blocks. The Gross Motor subtest consists of 72 items which measure head and trunk control and movements, and motor planning (Albers & Grieve, 2007; Case-Smith & Alexander, 2010).

The Social–Emotional Scale evaluates the social and emotional development of young children through 35 items which aim to assess children's emotional signals and gestures, such as smiling, cooing or reaching out for a hug, in their everyday life. Thus, the scoring of the Scale relies solely on the information provided by a child's parent or primary caregiver, since the examiner is interested in knowing what the child usually does, and not what a child is capable of doing (Albers & Grieve, 2007; Breinbauer et al., 2010).

Finally, the Adaptive Behaviour Scale utilises the Parent / Primary Caregiver Form of the Adaptive Behaviour Assessment Scale – Second Edition (ABAS–II), which consists of 241 items (Oakland, 2011). It assesses the practical, social, and conceptual skills, which allow a child to adapt to the demands of everyday life, such as communicating basic needs, following rules and/or directions, crawling or walking to get to desired locations, toileting, washing hands or brushing teeth, using manners, getting along with others and recognizing emotions. This scale is completed by the child's parent or primary caregiver (Harman & Smith-Bonahue, 2010).

For each of the five scales, a composite score with a mean of 100 and a standard deviation of 15 is calculated, indicative of a child's developmental status relative to the normative sample. Children with scores of one standard deviation below the mean in any of the five key domains are considered to have moderate–to–severe

neurodevelopmental delay (S. Johnson et al., 2014). In this study, we assessed language outcome at 2 years CGA in children born preterm using the Language Scale of the Bayley–III (see Chapter 6). Previous research has shown that the Cognitive and Language Scales of the Bayley–III can reliably predict long–term intellectual delay and language disorder in children born preterm (Bode et al., 2014; Torras-Mañá et al., 2014).

3.6. Brain MRI acquisition

sMRI and dMRI scans were obtained at term–equivalent age (38 - 42 weeks GA). Infants underwent brain MRI scans without sedation, having been fed, swaddled, and allowed to sleep naturally in the scanner (Boardman et al., 2020). All neonates were provided with flexible earplugs and neonatal earmuffs (MiniMuffs, Natus) for hearing protection. Vital signs were monitored throughout the scan, and all procedures were supervised by a physician, or a paediatric nurse trained in neonatal resuscitation.

For structural imaging, a Siemens MAGNETOM Prisma 3–Tesla MRI clinical scanner (Siemens Healthcare, Erlangen, Germany) and a 16–channel phased–array paediatric head receive coil were used to acquire a 3–dimensional T2–weighted sampling sequence with application–optimised contrasts by using flip angle evolution (SPACE) structural scanning (voxel size = 1 mm isotropic) with echo time = 409 ms and repetition time = 3200 ms.

For dMRI, a Siemens MAGNETOM Verio 3–Tesla MRI clinical scanner (Siemens Healthcare, Erlangen, Germany) and a 12–channel phased–array paediatric head

receive coil were used. The dMRI data consisted of 11 T2–weighted and 64 diffusion–weighted (b = 750 s/mm^2) single–shot, spin–echo, echo planar imaging volumes collected in the axial plane with 2 mm isotropic voxels (repetition time = 7300 ms, echo time = 06 ms, field of view = 256 mm, acquired matrix = 128 × 128, 50 contiguous interleaved slices with 2 mm thickness, acquisition time = 9 min 29 s).

3.7. Statement of consent

Written informed consent from parents/carers was obtained for all neonates.

3.8. Ethical approval

Ethical approval was obtained from the UK National Research Ethics Service (NRES), South East Scotland Research Ethics Committee (NRES numbers 11/55/0061 and 13/SS/0143 (phase I) and 16/SS/0154 (phase II)), and NHS Lothian Research and Development (2016/0255).

CHAPTER 4: METHODOLOGY FOR DATA ANALYSIS

4.1. Conventional mathematical notation

In the succeeding sections, the following conventional mathematical notation is used: lower-case letters represent scalars (e.g., x), bold lower-case letters represent vectors (e.g., $\mathbf{x}$ and $\mathbf{y}$), bold capital letters denote matrices (e.g., $\mathbf{X}$), and capital letters in italics denote random variables (e.g., X). The subscript i in the form $\mathbf{x}_i$ indicates the i^{th} element of a vector, and the subscripts ij in the form $\mathbf{X}_{ij}$ indicate the i^{th} observation of the j^{th} feature of a matrix $\mathbf{X}$. We will use n to represent the number of observations in a given dataset, and p to indicate the number of explanatory variables.

4.2. Introduction

The term *statistical learning* refers to a set of techniques aimed at understanding data, exploring relationships between variables, and accurately predicting an outcome for future observations. Statistical learning may be *supervised* or *unsupervised*. Supervised statistical learning tools are used to estimate a response based on one or more explanatory variables. On the other hand, when the response is unknown, unsupervised learning techniques can be used to infer relationships between the variables or between the observations and understand the structure of the data (James et al., 2013). In this study, we focus primarily on supervised learning.

Suppose that we have p explanatory variables, $X_1, X_2, X_3, \ldots, X_p$, each comprising n samples. Statistical learning tools aim at estimating the function

$$f(\mathbf{X}) = \mathbf{y} \tag{3.1}$$

where $\mathbf{y}$ is the response variable (i.e., a column vector) and $\mathbf{X}$ is the design matrix comprising all p explanatory variables, $X_1, X_2, X_3, \ldots, X_p$. The explanatory variables are also known as *input variables, independent variables*, *features*, or *predictors*. The response variable $\mathbf{y}$ is also known as the *output variable*, *outcome measurement* (or simply *outcome*), or *dependent variable*. To avoid any confusion, in the remaining sections, we will use the terms *feature* and *response* to refer to the explanatory variables $X_1, X_2, X_3, \ldots, X_p$ and $\mathbf{y}$, respectively. The function *f* is the *prediction model* or *learner* and aims to characterise the underlying relationship between the features summarised in the design matrix $\mathbf{X}$ and the response $\mathbf{y}$ and is generally unknown. Our goal is to determine the unknown function *f* by applying a statistical learning method to a training set (i.e., a dataset where both the features and the response are known), while a testing set (i.e., a dataset where only the features are known) is subsequently used to evaluate the performance of the prediction model. The statistical learning tools used for this task can be broadly placed into two distinct categories: *parametric*, and *non-parametric* (James et al., 2013). Parametric methods simplify the problem of estimating an entirely arbitrary function *f* by imposing a pre-specified structure on the functional form of *f*, thus limiting the problem of estimating *f* down to one of estimating a fixed set of parameters. Non-parametric methods do not impose any pre-specified structure and allow the data itself to determine the structural form of *f*. If the response is continuous, determining *f* is known as a *regression* problem. If the response is discrete, determining *f* is referred to as a *classification* problem (James et al., 2013).

This chapter presents the methodology we have used for data analysis in supervised regression and classification settings and comprises the following core steps: (a) exploratory data analysis, (b) dimensionality reduction techniques, (c) statistical mapping, (d) model evaluation, and (e) model validation. We will analyse each of these steps further in the following sections.

4.3. Exploratory data analysis

The initial step in data analysis is data exploration in order to understand the data structure. We begin by visualizing the data and performing correlation analysis in order to quantify any statistical associations between the features and the response, as well as between the features.

4.3.1. Data visualisation

We begin our exploratory data analysis by producing graphs, which allow to identify outliers, trends, and patterns in the data. The type of plots we use depends on the type of the variables. We use *bar charts* to visualise nominal and ordinal data. For continuous variables, *histograms* and *density plots* can be used. *Boxplots* are used to plot both discrete and continuous variables. In addition to these plots, we draw *scatterplots* in order to visualise potential relationships between each feature and the response, or between the features.

4.3.2. Correlation analysis

Scatterplots facilitate the visual assessment of the strength of the association between two variables. However, they do not provide us with a numeric measure of the strength of the association. In order to quantify the strength of the association between two variables, one approach is to use correlation coefficients. There are several types of correlation coefficients, but the *Pearson product-moment correlation coefficient*, and the *Spearman's rank correlation coefficient* are the most commonly used. The Pearson correlation coefficient (r) (Benesty et al., 2009) is used to quantify the strength of the linear association between two numeric variables and is defined as the ratio between the covariance of two random variables and the square root of the product of their variances:

$$r = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{3.2}$$

where X and Y are two random variables, cov is the covariance, $var(X)$ is the variance of X, and $var(Y)$ is the variance of Y.

If the variables do not follow a normal distribution or either or both of the variables are ordinal, we use the Spearman's rank correlation coefficient (r_s), which is effective in quantifying general monotonic relationships (Dodge, 2008; Spearman, 1904). The Spearman's rank correlation coefficient is defined as:

$$r_s = 1 - \frac{6\Sigma_{i=1}^{n} \mathrm{d}_i^2}{n(n^2-1)} \tag{3.3}$$

where d_i is the difference between the ranked values of two random variables for the i^{th} observation, $i \in \{1, \dots, n\}$, and n is the total number of observations.

The *phi-coefficient (φ)* is used to quantify the strength of the association between two dichotomous variables (Allen, 2017; Kraska-MIller, 2013) and is defined as the square root of the ratio between the chi-square and the sample size:

$$\varphi = \sqrt{\frac{x^2}{n}} \tag{3.4}$$

where x^2 is the computed chi-square statistic, and n is the number of observations. For a matrix of n x p size, the chi-square statistic is calculated by the formula:

$$x_c^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(\mathrm{O}_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}} \tag{3.5}$$

where c is the degrees of freedom defined as

$$df = (n-1)(p-1), \tag{3.6}$$

O_{ij} corresponds to the observed cell frequencies, and E_{ij} corresponds to the expected cell frequencies which we would be observed if some null hypothesis is true (for more details about hypothesis testing, see section 4.9.2) and is given by the formula:

$$\mathrm{E}_{ij} = \frac{R_i \mathrm{C}_j}{n} \tag{3.7}$$

where R_i is the i^{th} row marginal total, and C_j is the j^{th} column marginal total.

The *point-biserial correlation coefficient (r_{pb})* is used to quantify the association between a continuous and a dichotomous variable (Kornbrot, 2014). Let us assume that we have a continuous variable X and a dichotomous variable Y which takes on the values “0” (group “0”) and “1” (group “1”). Then, the point-biserial correlation coefficient is defined as:

$$r_{pb} = \frac{\mu_1 - \mu_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{3.8}$$

where μ_1 is the mean of the continuous variable X for group “1”, μ_0 is the mean of the continuous variable X for group “0”, s_n is the standard deviation of the continuous variable X for all observations, n_1 is the number of observations in group “1”, n_0 is the number of observations in group “0”, and n is the total number of observations.

The possible values of all the aforementioned correlation coefficients range from -1 to +1. The sign of the correlation coefficient denotes a positive or negative association between the two variables, and the magnitude corresponds to the strength of the correlation. The higher the magnitude, the stronger the statistical relationship between two variables is. Values close to +1 indicate a near perfect positive correlation, values close to zero indicate that there is no association between the two variables, and values close to -1 indicate a near perfect negative association, meaning that a

decrease in the value of one variable is associated with an increase in the value of the other variable. There is no universal guideline to determine when a bivariate relationship is statistically strong, as it depends on the application (Cohen et al., 2002). In this study, we consider that an absolute value of a correlation coefficient > 0.3 corresponds to a statistically strong association, in accordance to similar studies in clinical contexts (Meyer et al., 2001; Tsanas et al., 2013).

4.4. Statistical hypothesis testing

The statistical hypothesis testing procedure comprises the following steps:

- The *null hypothesis* and the *alternative hypothesis* are defined. In this case, the null hypothesis indicates that there is no difference in the performance between the two models, whereas the alternative hypothesis states that there exists a difference between the two models.
- The level of significance, α, is determined. In this study, we have used the commonly accepted value of α=0.05.
- Identification of the appropriate statistical hypothesis test and calculation of the test statistic and associated *p-value*.
- Reject or fail to reject the null hypothesis. If the computed *p*-value is less than the pre-specified level of significance, there is sufficient evidence to reject the null hypothesis. In contrast, if the *p*-value is equal to or greater than the significance level, we conclude that there is insufficient evidence to reject the null hypothesis.
- Interpretation of the results.

In this section, we will briefly describe the statistical hypothesis tests we have used in this study.

4.4.1. Paired t – test

A paired t–test is used to assess whether the difference between the means of two paired measurements is equal to zero (Ross & Willson, 2017). The underlying assumption is that the differences between the pairs of measurements must be normally distributed. In the context of model validation, the null hypothesis for the paired t–test is that the mean difference of the errors of the two models (i.e., the trained and the naïve, or two trained machine learning models) equals zero. The alternative hypothesis is that the mean difference of the errors of the two models does not equal zero. The *t-statistic* of the paired t-test is given by the formula:

$$t = \frac{\bar{x}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n}\right)}} \tag{3.36}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i$ is the sample mean (mean difference) and $\hat{\sigma}^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{x}_i - \bar{x})^2}$ is the sample standard deviation (n denotes the total sample size). Under the null hypothesis, the t-statistic follows the t-distribution with n-1 degrees of freedom. The corresponding *p*-value for the t-statistic can be calculated as the cumulative probability $\Pr(T > |t|)$ for a variable T that follows a t-distribution with n-1 degrees of freedom. The means of the errors of the two models are considered to be statistically significantly different if the *p*-value is less than the chosen alpha value of 0.05.

4.4.2. Wilcoxon signed – rank test

The Wilcoxon signed-rank test is the non-parametric alternative test to the paired t–test (Wilcoxon, 1945), performed when the differences between the pairs of measurements do not follow the normal distribution. The null hypothesis for this test is that the median difference is equal to zero, and the alternative hypothesis is that the median difference is not equal to zero. To compute the test statistic, first, the differences between the paired data samples are computed, and then the absolute values of the difference scores are ranked in increasing order, excluding any difference scores that are equal to zero. When there are tied scores present in the data, the average of the ranks involved is assigned to all scores tied for a given rank. Next, the sum of the ranks of the positive differences and the sum of the ranks of the negative differences are calculated. Finally, the absolute value of the minimum of the two sums of ranks is defined as the test statistic. The null hypothesis is rejected if the associated *p*-value is less than 0.05.

4.5. Curse of dimensionality

A problem that often arises, in both regression and classification problems, when we analyse high dimensional data (i.e. datasets with a large number of features / dimensions) is the *curse of dimensionality* (Bellman, 1966); given that there is only a finite number of available samples, as the dimensions of the feature space increase, the data become sparse making it difficult to prove a result statistically significant (Hastie et al., 2009). Prediction performance can be improved by reducing the number of dimensions of the feature space, a process that is called *dimensionality reduction*. Dimensionality reduction techniques can be divided into two main categories: *feature*

transformation and *feature selection*. Feature transformation aims at finding a low-dimensional representation of the original dataset, while feature selection refers to the process by which only a subset of the features from the original feature space is selected to build a learning model (Guyon & Elisseeff, 2003). In this study, our aim is to build a statistical model which accurately predicts the response while selecting a small feature set, hence improving model interpretability, and potentially reducing the prediction error. For that purpose, we use a number of feature selection methods which we will describe in the following sections.

4.6. Feature selection methods

Feature selection aims at reducing dimensionality by retaining only a subset of the features from the original feature space while irrelevant or redundant features are discarded. Feature selection techniques can be broadly placed into three main categories: *wrapper*, *embedded* and *filter* methods, based on the way the feature selection search is combined with the development of the prediction model (Guyon & Elisseeff, 2003; Hira & Gillies, 2015; Saeys et al., 2007). Wrapper methods take a particular machine learning method into account in order to choose the best subset of the original features. They evaluate multiple models by training and testing in the feature space. This means that wrappers are computationally inefficient, although they often provide the best results and optimise the performance of the particular machine learning model that was used (Hira & Gillies, 2015; Saeys et al., 2007). Embedded methods attempt to determine the best performing subset of features while the learning model is being constructed (Hira & Gillies, 2015; Saeys et al., 2007). This means that the resulting feature subset is specific to a particular learning algorithm, as in wrapper methods. However, embedded techniques are computationally more efficient than wrapper methods. Filter techniques work independently of a predictive model operating only at the intrinsic properties of the data, and thus provide a more general subset of features not tuned for a specific learning algorithm, which can be a disadvantage. Filters assess features based on some criterion, such as correlation coefficients or other statistical properties, and once the best performing feature subset has been determined, different prediction models can be evaluated. Filter methods are usually (but now always) computationally efficient and faster than both wrapper and embedded methods (Hira & Gillies, 2015; Saeys et al., 2007; Tsanas, 2022). In the

following sections, we present a summary of the feature selection algorithms we have used in this study.

4.6.1. Least Absolute Shrinkage and Selection Operator (LASSO)

The *Least Absolute Shrinkage and Selection Operator (LASSO)* (Tibshirani, 1996) is an example of an embedded feature selection technique. It is a shrinkage method that can be used for feature selection and shrinkage in linear regression models (Hastie et al., 2009); it shrinks the regression coefficients towards zero and thus performs feature selection. The LASSO uses the *L1 penalty*, which is equal to the sum of the absolute values of the coefficients, in order to regularise the magnitude of the coefficient estimates. Specifically, the LASSO coefficients $\beta_{0,}\beta_{1,}\dots,\beta_p$ are the values that minimise the quantity

$$\sum_{i=1}^{n}\left(\mathrm{y}_i-\beta_0-\sum_{j=1}^{p}\beta_j\mathbf{X}_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{3.9}$$

where n is the number of observations, p is the number of features, y_i is the value of the response for the i^{th} observation, β_0 is the intercept, β_j is the j^{th} regression coefficient, $\mathbf{X}_{ij}$ denotes the i^{th} value of the j^{th} feature ($j\in\{1,\dots,p\}$) belonging to the matrix $\mathbf{X}$, RSS is the Residual Sum of Squares* and $\lambda \geq 0$ is a regularisation parameter. Typically, we do not penalise the intercept β_0. The regularisation parameter λ in equation 3.9 controls the relative impact of the RSS and the L1 penalty on the regression coefficients. When λ is equal to zero, the L1 penalty has no effect, but as

* For definition of the RSS, see section 4.7.3.

the value of λ increases, the impact of the L1 penalty grows, forcing some of the coefficients to become exactly equal to zero (thus reducing the number of features), and at very large values of λ all coefficients will actually shrink to zero. As a result, the LASSO is able to yield sparse models (i.e. models comprising only a subset of features), serving as an embedded feature selection technique (James et al., 2013; Tibshirani, 1996).

4.6.2. Random forests feature importance

Random Forests (RF) feature importance is another example of an embedded feature selection method. When we use RF (see section 4.8.3) for regression and classification, it is possible to measure the importance or contribution of each feature in predicting the response. At each split in each tree, the improvement (i.e., the amount of reduction) in the split – criterion (i.e., the RSS[*] in regression settings, or the Gini index in classification settings) is the importance measure attributed to the splitting variable and is accumulated over all the trees in the forest for each feature. Thus, we can obtain an overall summary of the importance of each feature. In the case of regression models, we measure the total amount that the RSS is decreased due to splits over a given feature, averaged over all trees. A large value denotes an important feature. Similarly, in the context of classification models, we can measure the total amount that the Gini index is decreased by splits over a given feature, averaged over all trees. Again, a large value indicates a predictive feature within this statistical learning mechanism (Hastie et al., 2009; James et al., 2013).

[*] See section 4.8.3.

4.6.3. Boruta

The Boruta algorithm is a wrapper feature selection technique built around the RF learner and works for both regression and classification problems. Boruta uses *permutation importance* calculated by RF to evaluate the predictive importance of the features. In this case, RF make use of the *out-of-bag* observations (i.e., the observations not used – due to bootstrap sampling – to fit a given decision tree when constructing a RF model). Suppose that we have p features, $X_1, X_2, X_3, \dots, X_p$, each comprising n samples. When a decision tree is grown, the out-of-bag observations are passed down the tree and the out-of-bag prediction error is obtained. Subsequently, the values of the feature X_j ($j \in \{1, \dots, p\}$) are randomly permuted in the out-of-bag observations, and these shuffled feature values are again passed down the tree and the resulting prediction error is recorded. The increase of the model's prediction error (i.e., the difference between the original out-of-bag prediction error and the prediction error obtained after random permutation) as a result of shuffling the values of the X_j feature is averaged over all trees, and corresponds to the importance attributed to the X_j feature in the RF. The standard deviation of the prediction error is also calculated. The Boruta algorithm uses *Z score* as the importance measure. In other words, it measures the importance of each feature by dividing the mean increase of the model's prediction error attributable to the feature by the standard deviation of the prediction error. However, the Z score is not directly related to the statistical significance of the feature importance returned by the RF learner since its distribution is not $\mathcal{N}(0,1)$. Hence, the Boruta algorithm solves this problem by creating copies which are called *shadow features* for each original feature of the feature set, whose values are obtained by random permutation of the values of the original features so as to remove their

correlations with the response variable. Subsequently, it builds a RF prediction model using all features (i.e., the original features and the shadow features) and evaluates the importance of all the features. After that, the Z score for each feature is calculated and then the *Maximum Z score among the Shadow Attributes (MZSA)* is identified. Among the original features, the ones which have importance significantly higher than the MZSA are deemed as important, and the rest are considered redundant and are permanently removed from the procedure. The process is repeated for a prespecified number of iterations (i.e., prespecified number of trees). The shadow features are subsequently discarded and the final ranked order of the original features is computed based on their Z scores (Kursa & Rudnicki, 2010). In summary, the Boruta algorithm comprises the following steps:

1. Extension of the original dataset by creating copies (shadow features) for each original feature.
2. Random permutation of the values of shadow features so that correlations with the response are removed.
3. The datasets of the original and shadow features are merged.
4. Training of a RF learner using the extended dataset and calculation of the Z scores for all features (i.e., original and shadow).
5. The MZSA is computed.
6. The original features which have importance significantly higher than MZSA are deemed as important.
7. The original features which have importance significantly lower than MZSA are eliminated.
8. All shadow features are discarded.

9. The above steps are repeated until importance is assigned to all original features, or the algorithm has reached a prespecified number of RF iterations.

4.6.4. ReliefF and Regressional ReliefF

Relief is a filter feature selection method which was developed by Kira and Rendell in 1992 specifically for binary classification problems including categorical and/or numeric features (Kira & Rendell, 1992). Later, in 1994, Kononenko proposed an extension of the Relief algorithm, ReliefF (Kononenko, 1994; Robnik-Šikonja & Kononenko, 2003), a more robust algorithm which can also be applied to multi-class classification problems and can deal with incomplete and noisy data. In 1997, ReliefF was further adapted to deal with regression problems and was called Regressional ReliefF (RReliefF) (Robnik-Sikonja & Kononenko, 1997; Robnik-Šikonja & Kononenko, 2003). The basic idea of the Relief algorithm is to assign a *weight* value to all features of a dataset based on how well their values distinguish between the samples that are near to each other and thus, how useful they are in predicting the response variable. The important features will have a large weight, while the redundant ones will have a low weight. For each feature vector belonging to one random observation, Relief searches for its two nearest (by Euclidean distance) feature vectors; the closest observation that belongs to the same class is called “nearest hit”, while the closest observation belonging to a different class is called “nearest miss”. A feature will be assigned a large weight value if it differs from its nearest neighbour of a different class more than it does from its neighbour of the same class. Similarly, a feature will be assigned a low weight value if it differs from its nearest neighbour of the same class more than it differs from its neighbour of a different class. The process is

repeated for m times, where m is a parameter determined by the user. Starting with a p-dimensional weight vector of zeros, which is updated in each iteration (m total iterations), the weight assigned to a feature X ($W[\mathrm{X}]$) is defined as:

$$W[\mathrm{X}] := W[\mathrm{X}] - \frac{diff(\mathrm{X}, \mathbf{r}_\mathrm{i}, \mathbf{h})}{m} + \frac{diff(\mathrm{X}, \mathbf{r}_\mathrm{i}, \mathbf{m})}{m} \tag{3.10}$$

where $\mathbf{r}_\mathrm{i}$ is a randomly selected sample represented by a vector of p feature values for $i \in \{1, \ldots, m\}$, **h** is the nearest hit (a vector of p feature values), **m** is the nearest miss (a vector of p feature values), and m is a user-defined parameter. The function *diff* calculates the difference between the values of a feature for two samples; let us assume we have two random samples **a**, and **b**, where each is represented by a p-dimensional vector, then the difference between the values of a feature X is defined as:

if X is categorical:

$$diff(\mathrm{a_x}, \mathrm{b_x}) = \begin{cases} 0; \mathrm{a_x} \text{ and } \mathrm{b_x} \text{ are equal} \\ 1; \mathrm{a_x} \text{ and } \mathrm{b_x} \text{ are different} \end{cases} \tag{3.11}$$

if X is numeric:

$$diff(\mathrm{a_x}, \mathrm{b_x}) = \frac{|\mathrm{a_x} - \mathrm{b_x}|}{\max(\mathrm{X}) - \min(\mathrm{X})} \tag{3.12}$$

where $\mathrm{a_x}$ is the value of feature X for **a** and $\mathrm{b_x}$ is the value of feature X for **b**.

The main difference between the original Relief and its extension ReliefF is that the latter searches for k nearest hits and misses and averages their contribution to the weight of each feature. Selection of k nearest hits and misses, instead of one nearest hit and miss, renders the algorithm less sensitive to noisy data. The parameter k is defined by the user, but usually a value of k = 10 works well for most applications (Kononenko, 1994; Robnik-Šikonja & Kononenko, 2003). In regression, nearest hits and misses cannot be used, but similar to ReliefF, for a random instance, RReliefF searches for its k nearest neighbours and assigns a large weight to the features whose values are different between instances with different response values, and a low weight to the features whose values are different between instances with the same response values. There are various feature evaluation measures that ReliefF and RReliefF can use, and which can be determined by the user. In this study, we will explore three different feature evaluation measures; "equalK" where k nearest instances have equal weight, "expRank" where k nearest instances have weight exponentially decreasing with increasing rank (rank of nearest instance is determined by the increasing distance from the selected instance), and "bestK" where all possible k nearest instances are tested and for each feature the highest score is returned (nearest instances have equal weights).

4.7. Voting scheme: defining the final feature subset

First of all, we need to note that it would be wrong to determine the final feature subset on the basis of all of the samples, and then use these features to test the performance of the prediction model on a test set, using for example cross-validation (see section 4.9.1), since the selected features will "have already seen" the test samples and hence

this will result in data leakage. Instead, the subset of the best features should be determined using cross-validation; the test samples must be “left out” before any feature selection algorithm is applied to the dataset (Hastie et al., 2009).

The feature selection algorithms described in the preceding sections, except for the LASSO, aim at ranking the features of a given dataset based on their contribution towards prediction of the response variable. Ideally, when using cross-validation, at the end of each iteration for any given feature selection algorithm, we would obtain the same order of ranked features which would clearly indicate which feature subset should be selected to train the machine learning model. However, in practice, the order of ranked features may be different at the end of each iteration for any given feature selection algorithm. Thus, in order to select the final feature subset, we follow the process described by Tsanas et al. in 2012 (Tsanas et al., 2012) and further refined and extensively used for feature selection algorithmic comparisons in Tsanas 2022 (Tsanas, 2022).

Specifically, for a given feature selection algorithm, when using cross-validation, at the end of each iteration, we obtain a vector of the ordered sequence of the indices of the features, where the first feature is considered to be the most important one, and the last feature corresponds to the least important one. We store these vectors in a matrix of n x p size. This way, in each of the rows of the matrix we have stored the feature subset selected at the end of each iteration. For example, the feature index in cell [1,1] corresponds to the most important feature generated from the first iteration, and the cell [17,4] contains the index of the fourth best feature generated from the seventeenth iteration. Subsequently, we apply the following voting scheme in order to obtain the

final feature subset for a given feature selection algorithm. First, we need to identify the feature index which has most frequently been ranked as first (i.e., most important) across all iterations, then we need to identify which feature appears most frequently as second or third and so on. In case a feature index has already been included in the final subset and is later found again as most frequent, we need to select the second most frequent and so on. Ties are resolved by including the lowest index number. The LASSO, however, does not rank the features from the most important to the least important one, but it may remove features in subsequent steps during its incremental feature selection search. Therefore, for LASSO, we need to obtain its regularisation path (i.e., the coefficient estimates for different values of the regularisation parameter λ) before we proceed with the voting scheme we have described above.

In the end, we obtain a vector which consists of feature indices ordered from the most important to the least important one, for a given feature selection algorithm. These features are then used to train the learner one by one (i.e., we first train the model with the most important feature, then with the two top – ranked features etc.). Therefore, we can calculate the performance of the model as a function of the number of features used, which are progressively inserted into the statistical learner. Following the *principle of parsimony*, we choose to keep the statistical model which has the greatest predictive power with the fewest possible features.

4.8. Statistical mapping

As we have previously discussed, statistical learning tools are divided into two main groups: parametric, and non-parametric. The main advantage of parametric methods, which impose a pre-specified structure on the functional form of f, is that they are generally easy to interpret. However, by imposing a functional form structure *a priori* that may be too far from the true form of f, will potentially lead to false interpretation of the properties of the data. On the other hand, non-parametric methods allow the data to determine the structure of the form of f (James et al., 2013). By avoiding imposing any pre-specified structure, non-parametric methods can fit a wider range of possible forms for f, compared to parametric methods. Nevertheless, non-parametric methods typically require far more data samples than parametric approaches in order to accurately estimate f (Breiman, 2001; James et al., 2013). In the following sections, we will present a summary of the parametric and non-parametric methods we have used in this study.

4.8.1. Multivariate linear regression

Linear regression is a parametric machine learning algorithm used to model the relationship between a continuous response and a number of features which may be continuous, discrete, or categorical. The linear regression model assumes that the true relationship between the features and the response is approximately linear or that the linear model is a reasonable approximation for the data. Linear regression reduces the problem of estimating an entirely arbitrary function f down to one of estimating a finite number of *parameters* or *coefficients* (Hastie et al., 2009; James et al., 2013; A. Schneider et al., 2010). The formula of the multiple linear regression model is:

$$f(\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \tag{3.13}$$

where $\mathbf{X}$ is a design matrix comprising all p features $X_1, X_2, X_3, \ldots, X_p$, X_j is the j^{th} feature ($j \in \{1, \ldots, p\}$, β_0 is the intercept of the linear model, and β_j denotes the slope terms of the linear model. The intercept and the slope terms together are the model parameters or coefficients and are generally unknown. In the linear regression formula, ε is a random error term which is assumed to be independent of the features in $\mathbf{X}$, with mean equal to zero and constant variance σ^2. The error term ε is a value which indicates what we may be missing with the linear regression model (e.g., measurement errors, unknown variables that influence the response).

We need to note that the linear regression model is linear in the coefficients. In order to estimate the regression coefficients, we need to fit the linear regression model on the available training set. The goal is to obtain coefficient estimates such that the linear model fits the available data well, meaning that the resulting line should be as close as possible to the observed data points. There are a number of methods for measuring closeness. The most common approach in estimating the coefficients involves minimizing the least squares criterion (James et al., 2013). Let us assume that y_i is the observed response value for the i^{th} observation, and $\hat{\mathrm{y}}_i$ is the estimated response value from the linear model for the i^{th} observation, $i \in \{1, \ldots, n\}$, n being the total number of observations. Then $\mathrm{e}_i = \mathrm{y}_i - \hat{\mathrm{y}}_i$ represents the i^{th} residual (or error). The residuals are exactly the vertical distance between the observed data point and the associated point on the regression line. Thus, the least squares approach estimates

the coefficients $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ using the values that minimise the RSS, which is defined as:

$$RSS = \sum_{i=1}^{n} \mathrm{e}_i^2$$

$$= \sum_{i=1}^{n} (\mathrm{y}_i - f(\mathbf{x}_i))^2$$

$$= \sum_{i=1}^{n} (\mathrm{y}_i - \beta_0 - \sum_{j=1}^{p} \mathbf{X}_{ij}\beta_j)^2 \qquad (3.14)$$

An alternative approach for fitting a linear regression model is the LASSO, which we have described in section 4.5.1.

Before fitting a linear model to a training set, there are five basic assumptions that need to be met (Poole & O'Farrell, 1971):

- Independence: the residuals of the linear regression model are assumed to be independent and identically distributed.
- Linearity: there is a linear relationship between the response and the features.
- Normality: the residuals of the model follow the normal distribution.
- Homoscedasticity: the residuals have a constant variance σ^2; the variances of the residuals are constant along the values of the response.
- Absence of multicollinearity: the features in a linear regression model should not be correlated with each other.

4.8.2. Multivariate logistic regression

Logistic regression is a parametric method used in classification settings to describe the relationship between a discrete response and one or more features which may be quantitative or qualitative. Logistic regression can be used for binary and multi-class classification problems, but in practice, it is mainly used for binary outcomes. For simplification purposes, we will assume that the response $\mathbf{y}$ is binary, taking the generic 0/1 coding. The key concept is that logistic regression models the *probability* that the response belongs to a particular category as a function of the p features of a matrix $\mathbf{X}$ of size n x p, and this constitutes the basis for making the classification (Hastie et al., 2009; James et al., 2013). We model the probability that y_i = 1, $i \in \{1,..,n\}$, that is $p(\mathbf{X})$ = Pr(y=1| $\mathbf{X}$), using the logistic function:

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{3.15}$$

The *odds* that y_i equals 1 is given by the formula:

$$\frac{p(\mathbf{X})}{1-p(\mathbf{X})} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p} \tag{3.16}$$

If we take the logarithm of both sides, we get the *log-dds* or *logit*:

$$log\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.17}$$

we see that the logistic regression model has a logit that is linear in the features $X_1, X_2, X_3, \ldots, X_p$. The β_0, β_1, …, β_p are the regression coefficients. The most common approach in estimating the coefficients is *maximum likelihood estimation*. Maximum likelihood seeks estimates for the coefficients such that the predicted probabilities that the response equals one of two categories correspond as closely as possible to the true category. In other words, maximum likelihood estimates the coefficients β_0, β_1, …, β_p using the values that maximise the *likelihood function*:

$$l(\boldsymbol{\beta}) = \prod_{i:y_i=1} p(\mathbf{x}_i) \prod_{i':y_{i'}=0} (1 - p(\mathbf{x}_{i'})) \tag{3.18}$$

where $\boldsymbol{\beta}$ is a vector comprising the regression coefficients β_0, β_1, …, β_p, and $\mathbf{x}_i$ is the feature vector for the i^{th} observation. The maximum log-likelihood for a binary response is given by the equation:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ \mathrm{y}_i \log\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) + (1 - \mathrm{y}_i) \log\left(1 - p(\mathbf{X})\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ \mathrm{y}_i \left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) - \log\left(1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}\right) \right\} \end{aligned} \tag{3.19}$$

Before conducting logistic regression, there are some key assumptions that need to be met (Tabachnick & Fidell, 2007):

- Independence: the residuals of the logistic regression model are assumed to be independent of each other.

- Linearity: logistic regression assumes linearity between the continuous features of the model and the logit of the response.
- Absence of multicollinearity: the features in a logistic regression model should not be correlated with each other.
- Absence of extreme outliers in the data.

In this study, we have used the *L1 regularised logistic regression* which uses the L1 penalty which is equal to the sum of the absolute values of the coefficients, in order to regularise the magnitude of the coefficient estimates and yield sparse models (Hastie et al., 2009). For L1 regularised logistic regression, the coefficients $\beta_{0,}\beta_{1,}\dots,\beta_p$ maximise the term:

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{n}\left\{\mathrm{y}_i\left(\beta_0+\beta_1X_1+\cdots+\beta_pX_p\right)-\log\left(1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}\right)\right\}-\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (3.20)$$

where n is the total number of observations, p is the number of features, β_0 is the intercept, β denotes the slope coefficients, β_j is the j^{th} coefficient, and $\lambda \geq 0$ is a regularisation parameter. As with LASSO (see section 4.6.1), we do not penalise the intercept term.

4.8.3. Tree – based learning algorithms

Tree – based learning algorithms are non-parametric techniques which can be used in both regression and classification settings in order to model the non-linear relationship between the features and the response of a given dataset. The main idea

of tree-based methods is to partition the feature space into a set of distinct hyper-rectangular regions. A training set is used to train the learner and subsequently, in order to make a prediction for a previously unseen observation (i.e., an observation that belongs to a validation set or test set) we use the mean (for continuous responses) of the response values or the most commonly occurring class (for discrete responses) of the training observations in the region to which it belongs (Hastie et al., 2009; James et al., 2013). First, we will briefly describe how *decision trees* work, and then we will expand on how these trees can be aggregated to construct more powerful learners.

4.8.3.1. Regression trees

We will begin by describing the process of growing a *regression tree*. Let us consider a dataset with p features $X_1, X_2, X_3, \ldots, X_p$, a continuous response $\mathbf{y}$, and n observations. Regression trees comprise a set of splitting rules which partition the feature space into J hyper-rectangular regions, known as *terminal nodes* or *leaves of the tree*. The algorithm needs to decide on the splitting features and split points in order to generate terminal nodes, or regions, R_1, …, R_J such that the RSS is as small as possible. The RSS is given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(\mathrm{y}_i - \hat{\mathrm{y}}_{R_j} \right)^2 \tag{3.21}$$

where R_j denotes the j^{th} region, and $\hat{y}_{R_j}$ is the mean of the response values of the training observations in the j^{th} region, $j \in \{1, 2, \ldots, J\}$. The computational requirements of searching through every possible combination of splitting features and split points

in terms of minimizing the RSS would be prohibitive. Thus, we proceed with a greedy algorithm, known as *recursive binary splitting*. We begin with all of the data belonging to one region (at the top of the tree). The algorithm goes through all the features of the dataset and through all possible splitting points s and decides on a splitting feature X_j and the best split point s such that dividing the feature space into two regions

$$R_1(j,s) = \{\mathbf{X} \mid \mathrm{X}_j \leq s\} \text{ and } R_2(j,s) = \{\mathbf{X} \mid \mathrm{X}_j > s\} \tag{3.22}$$

minimises the RSS of the resulting tree:

$$\sum_{i:x_i \epsilon R_1(j,s)} \left(\mathrm{y}_i - \hat{\mathrm{y}}_{R_1}\right)^2 + \sum_{i:x_i \epsilon R_2(j,s)} \left(\mathrm{y}_i - \hat{\mathrm{y}}_{R_2}\right)^2 \tag{3.23}$$

where $R_1(j,s)$ is the region of feature space in which the feature $\mathrm{X}_j$takes on a value less than or equal to *s*, and $\hat{\mathrm{y}}_{R_1}$is the mean response of the training observations in $R_1(j,s)$, $R_2(j,s)$ is the region of feature space in which the feature X_j takes on a value greater than *s*, and $\hat{y}_{R_2}$is the mean response of the training observations in $R_2(j,s)$. Subsequently, for each of the resulting regions, the process is repeated until each terminal node contains some minimum number of observations, which is typically set to 5 or 10, resulting in a large tree T_0. Then, in order to make a prediction for a given test observation, we use the mean response of the training observations in the region to which the test observation belongs.

The process described yields large regression trees that might overfit the training data and fail to generalise to test or validation data. To prevent overfitting, the resulting tree

T_0 needs to be pruned, meaning that some of its internal nodes (i.e., the points where the feature space is partitioned) need to be collapsed in order to obtain a subtree $T \subset T_0$ which will generalise better to previously unseen data. This can be achieved by applying *cost complexity pruning* (Hastie et al., 2009; James et al., 2013) to the full sized tree T_0 in order to obtain a sequence of subtrees (from the full sized tree T_0 to a single-node tree) as a function of a tuning parameter *α*. For each value of the tuning parameter *α*, there is a unique subtree $T \subseteq T_0$ that minimises the *cost complexity criterion*:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(\mathrm{y}_i - \hat{\mathrm{y}}_{R_m}\right)^2 + a|T| \qquad (3.24)$$

where m denotes terminal notes, $|T|$ is the number of terminal nodes of the subtree T, R_m is the m^{th} region, and $\hat{\mathrm{y}}_{R_m}$ is the predicted response in the m^{th} region, and a is a nonnegative tuning parameter. It is easy to see that when a is equal to zero, then the subtree T is equal to the initial large tree T_0. As we increase a, the tree T_0 gets pruned. We estimate the value of a using cross-validation (see section 4.10.1) and obtain the subtree that corresponds to the chosen value of a.

4.8.3.2. Classification trees

The process of building a classification tree is very similar to the one described for regression trees with the only changes pertaining to the criteria for making the binary splits and pruning the resulting tree. For a discrete response taking values 1, 2, …, K,

instead of using the RSS as the splitting criterion, we may use the *Gini index*, which is given by the formula:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.25}$$

where $\hat{p}_{mk}$ denotes the proportion of training observations in the m^{th} region that are from the k^{th} class. The Gini index is a measure of node purity; a small value indicates that a node contains mainly observations from the same class. Alternatively, another measure of node purity that can be used as splitting criterion is *cross-entropy*, which is defined as:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{3.26}$$

Small values of cross-entropy indicate that a node is pure. When pruning a classification tree, the *classification error rate* is used (although the Gini index or cross-entropy can also be used) and is given by the equation:

$$E = 1 - \max_{k}(\hat{p}_{mk}) \tag{3.27}$$

Subsequently, in order to make a prediction for a test observation, we simply use the most commonly occurring class of the training observations in the region to which the test observation belongs.

Decision trees are very easy to interpret, mimicking clinicians' decision making. However, they suffer from high variance, meaning that even small changes in the training data may completely alter the final estimated tree. For this reason, ensemble of trees, such as *random forests*, which have low variance, are used to produce more powerful learners (Hastie et al., 2009; James et al., 2013).

4.8.3.3. Random forests

RF (Breiman, 2001) involve fitting decision trees on bootstrapped samples of the original training data and then combining all individual trees to create a single powerful predictive model. The trees are grown fully and there is no pruning. The process of building an individual tree is similar to the one described so far, but in RF, each time a split in a tree is considered, a random sample of m features is chosen as split candidates from the full set of p features. Each split is allowed to use only one of those m features. The number of randomly selected features to be considered for each split when growing a tree is the only hyperparameter typically used for optimizing RF performance provided the number of trees is set to be sufficiently large (in this study, we have set the number of trees to be equal to 500). Typically, when building a RF of classification trees, m is set to be equal to the square root of the number of features in the training data, and in regression settings, m is set to be equal to a third of the number of features in the training data (Hastie et al., 2009). In this study, we have optimised the number of features randomly selected as split candidates; starting with the typical values of $m = \sqrt{p}$ for classification, and $m = {}^{p}/_{3}$ for regression, we have performed hyperparameter tuning using the out-of-bag observations of the training set in order to get an estimate of the out-of-bag error as a function of the number of

randomly selected features. When used in regression, RF make predictions by averaging the resulting predictions of the decision trees used to build the learner. For classification problems, the learner obtains a class vote from each individual tree, and then classifies taking a majority vote: the overall prediction is the most commonly occurring class.

4.9. Model performance metrics

After fitting a statistical learning model on the training data, we need to evaluate its predictive performance. In other words, we need to quantify the extent to which the predictions of the learner for previously unseen data (i.e., data that was not used during the training of the model) are close to the true outcome values of the data. We use the *Mean Squared Error (MSE)*, *Root Mean Squared Error (RMSE), Mean Absolute Error (MAE),* and *Mean Absolute Percentage Error (MAPE)* to evaluate regression models, and *confusion matrices* (from which we calculate three quantities: *accuracy, sensitivity,* and *specificity*) for the evaluation of classifiers. We need to note that model performance is assessed using out-of-sample data, i.e., not the training data that was used to train the learner.

4.9.1. Evaluation of regression models

Regression models are usually evaluated using the MSE (James et al., 2013) which is the mean of the squared difference between the predicted values of a response and the true responses. Given a dataset with n observations, the MSE is given by the formula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2 \tag{3.28}$$

where y_i is the true value of the response for the i^{th} observation, and $\hat{\mathrm{y}}_i$ is the predicted value of the response for the i^{th} observation. Due to its quadratic nature, the MSE is sensitive to outliers and penalises large errors and thus, it is particularly useful when large errors are undesirable. Values can range from 0 to ∞, with zero denoting that the model fits the data perfectly (i.e., all predicted values of the responses are identical to the true values). Smaller values of the MSE indicate that the predicted responses are close to the true ones, meaning that the model fits the data well. The units of the MSE are squared units of the response which makes the metric difficult to interpret. Hence, we usually report the RMSE, which is simply the square root of the MSE:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2} \tag{3.29}$$

Due to the square root, the units of the RMSE are the same as the units of the response. We usually report the MAE of a model, as an additional, intuitively understandable performance measure. The MAE is defined as the mean of the absolute differences between the predicted values of the response and the true responses:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathrm{y}_i - \hat{\mathrm{y}}_i| \tag{3.30}$$

The values of the MAE can range from 0 to ∞, with smaller values indicating better performance of the prediction model. The MAE is more robust to outliers compared to the MSE and RMSE. We need to note that the RMSE is always greater or equal to the MAE. Another metric that is usually reported is the MAPE, which is the mean of the absolute percentage errors between the predicted and the true values of the response:

$$\mathrm{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|\mathrm{y}_i - \hat{\mathrm{y}}_i|}{\mathrm{y}_i}\right) \tag{3.31}$$

4.9.2. Evaluation of classification models

The accuracy of a classification model can be assessed by constructing a confusion matrix which is a contingency table of the observed and predicted classes (Stehman, 1997). Table 4.1 is an indicative example of a confusion matrix for binary classification problems, and it can be generalised for multiclass classification problems.

Table 4.1. Format of a confusion matrix for binary classification problems.

		Reference	
		Positive	**Negative**
Prediction	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

The columns of the confusion matrix are the observed classes, and the rows are the predicted classes. The long diagonal of the confusion matrix corresponds to the cases where the model is correct (true positive, true negative), while the other diagonal includes the cases where the model is incorrect (false positive, false negative). True positive indicates the number of positive cases that were correctly classified by the model, true negative refers to the number of negative cases that were correctly classified, false positive refers to the number of negative cases that were misclassified, and false negative denotes the number of positive cases that were wrongly classified by the model as negative. The evaluation of a classifier involves measurement of three metrics: *accuracy*, *sensitivity,* and *specificity*.

Accuracy is useful when the data is balanced and is defined as the ratio between the correctly classified observations and the total number of observations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.32}$$

However, if a dataset is highly imbalanced in terms of the response, and the prediction model classifies all cases to the most commonly occurring class, then accuracy will be high. In this case, we use an alternative metric: *balanced accuracy*, which is given by the formula:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{3.33}$$

Sensitivity is the ratio between the number of true positive predictions and the total number of observations which are actually positive for the outcome of interest (i.e., true positives and false negatives):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.34}$$

Specificity is defined as the ratio between the number of true negative predictions and the total number of observations which are actually negative for the outcome of interest (i.e., true negatives and false positives):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.35}$$

4.10. Model validation

The ultimate goal of a machine learning model is to accurately estimate the response values for previously unobserved data which have not been used in the training process. Hence, we need an additional dataset in order to assess model performance. However, in practice, an additional external validation dataset is often not available. Instead, we can use resampling techniques which involve repeatedly drawing a subset of samples from the original dataset to be used in training the model, while the

remaining samples serve as a test set to evaluate the performance of the fitted model (James et al., 2013). In this study, we have used *cross-validation* (James et al., 2013) in order to estimate the test error rate of the trained models.

4.10.1. Cross-validation

The key idea in cross-validation is to repeatedly hold out a subset of the original data (i.e., test set) from the training process, train the model using the remaining data (i.e., training set), and then apply the fitted model to those held out observations in order to assess its performance. It provides an estimate of the performance of the model on new unobserved data, given that the new data comes from the same distribution as the data that was used to train the model. In this study, we have used k*-fold cross-validation* and *leave-one-out cross-validation*.

In k-fold cross-validation, the original dataset is randomly split into k approximately equal-sized folds or groups and the statistical learning model is subsequently trained using the k-1 folds, while the fold that has been held out is used as a test set to get an estimate of the test error rate of the trained model. The process is repeated k times and each time a different fold is used as the test set. The result is k estimates of the test error. The final test error rate can be calculated by averaging these k resulting test error estimates. Determining the value of k is associated with a bias – variance trade-off (Hastie et al., 2009; James et al., 2013). Typically, we set k to be equal to 5 or 10, as these values offer a good bias-variance trade-off.

A special case of k-fold cross-validation is *leave-one-out cross-validation (LOOCV)* where k is set to be equal to the sample size n of the original dataset. LOOCV involves holding out a single observation to be used as the test set, while the learner is trained using the remaining n-1 observations. The process is repeated n times and each time a different observation from the original dataset is used as the test set. The result is n estimates of the test error. The final test error rate is the average of these n test error estimates. Compared to k-fold cross-validation with $k < n$, LOOCV is computationally more expensive, since it involves fitting the same model as many times as the number of observations of the original dataset. In addition, the test error estimates that are obtained from LOOCV have low bias but suffer from high variance as they are highly correlated with each other, since they come from models which have been fitted on subsets of the original data that are very similar to each other (Hastie et al., 2009; James et al., 2013). In this study, due to the limited sample size available, we have used LOOCV throughout.

4.10.2. Surrogate testing

In this study, we have used statistical hypothesis testing (see section 4.4) in order to compare the performance of two machine learning models, as well as to demonstrate whether the selected model significantly outperforms a naïve benchmark (i.e., a model that always predicts the mean or median value of the response) in the context of regression problems. There is no threshold for the MSE so to evaluate the performance of a model the MSE needs to be compared to a benchmark, usually the MSE of a naïve (surrogate) model that would always predict the mean of the response values. A paired t – test (Ross & Willson, 2017) was performed to compare the

performance metric (e.g., MSE) between two machine learning models, as well as between the trained model and a naïve (surrogate) model. The assumption of the normal distribution of the differences of the pairs was tested using a normal probability quantile-quantile plot. In case this assumption was violated, the non-parametric alternative Wilcoxon signed – rank test was used (Wilcoxon, 1945).

4.11. Data balancing techniques

In classification settings, a dataset is imbalanced when the proportion of classes is uneven. In real world data, we may come across a variety of imbalance ratios between the cases in the minority class and those that belong to the majority class (e.g., 1:10, 1:100, or even 1:1000). The problem that arises when dealing with imbalanced datasets is that there are too few cases in the minority class compared to the majority class, and thus, machine learning algorithms cannot be trained effectively and tend to considerably underestimate (or suppress) the minority class. However, it is the minority class that we are often more interested in, as the majority class usually includes the "normal" cases, while the minority class includes the "abnormal" cases which we want to predict. In addition, another issue is that when a learner is trained on an imbalanced dataset, the use of single traditional performance metrics, such as accuracy, may be misleading (see section 4.9.2). There are a number of data balancing techniques which aim to transform the training data in order to achieve an equal representation of the classification categories. These techniques can be broadly placed into three main categories: oversampling techniques, undersampling techniques, and a combination of oversampling and undersampling. There is no single best method, so in this study, we experimented with different data balancing techniques: *random oversampling* of

the minority class (He & Ma, 2013), *random undersampling* of the minority class (He & Ma, 2013), *Synthetic Minority Oversampling Technique (SMOTE)*, and a combination of oversampling and undersampling techniques (i.e., random oversampling or SMOTE paired with random undersampling) (Chawla et al., 2002). We need to note that all data balancing techniques need to be performed on the training set, which is used to train the learner, and not on the test set which is used to get an estimate of the performance of the model on unseen observations. In this section, we will present a summary of the techniques we have used in this study.

Random oversampling involves randomly selecting cases that belong to the minority class, with replacement, and adding them to the training dataset until a more balanced distribution of cases is achieved (He & Ma, 2013). Random undersampling works by randomly discarding cases from the majority class in the training set, with or without replacement, until there is an equal or almost equal number of cases between the minority and the majority class (He & Ma, 2013). SMOTE involves oversampling the minority class by generating new synthetic samples to create a balanced dataset (Chawla et al., 2002). Specifically, for each minority class sample, the k minority class nearest neighbours are identified by Euclidean distance. The value of k is typically set to be equal to 5. Depending upon the number of synthetic samples that need to be generated, a random number or all of the k nearest neighbours are used to calculate the difference between the feature vector under consideration (i.e., the minority class sample for which we want to identify the k nearest neighbours) and each of its nearest neighbours. Subsequently, the computed difference is added to the feature vector under consideration and as a result, a new synthetic sample is created along the line segments joining any or all of the k minority class nearest neighbours. It is easy to

understand that the aforementioned procedure can be applied to datasets which comprise only continuous features. Two extensions of SMOTE: *Synthetic Minority Oversampling Technique – Nominal Continuous (SMOTE-NC),* and *Synthetic Minority Oversampling Technique – Nominal (SMOTE-N)* have been developed to deal with datasets comprising both continuous and categorical features, or only categorical features, respectively (N. V. Chawla et al., 2002). In SMOTE-NC, identification of the k nearest neighbours involves adding the median of standard deviations of all continuous features for the minority class to the Euclidean distance equation, when the categorical features between feature vectors differ. Subsequently, the continuous features of the new synthetic sample are calculated using the procedure described for SMOTE, while the categorical features are given the most frequently occurring value among the identified k nearest neighbours. Finally, SMOTE-N deals with datasets which comprise exclusively categorical features. In this case, identification of the k nearest neighbours involves using a modified version of the *Value Difference Metric*, which computes the distance between values of categorical features (Cost & Salzberg, 1993). The distance δ between two values for a given feature is given by:

$$\delta(V_1, V_2) = \sum_{i=1}^{h} \left| \frac{C_{1_i}}{C_1} - \frac{C_{2_i}}{C_2} \right|^m \tag{3.37}$$

where V_1 and V_2 are two possible values for a given feature, C_{1_i} denotes the number of occurrences of feature value V_1 for class i, C_1 is the total number of occurrences of feature value V_1, C_{2_i} denotes the number of occurrences of feature value V_2 for class i, C_2 is the total number of occurrences of feature value V_2, h is the number of classes

of the outcome in the dataset, and m is a constant usually set to 1. Subsequently, the total distance Δ between two feature vectors is given by the formula:

$$\Delta(\mathbf{x}, \mathbf{z}) = w_x w_z \sum_{i=1}^{p} \delta(\mathrm{x}_i, \mathrm{z}_i)^r \tag{3.38}$$

where $r = 1$ gives the Manhattan distance, and $r = 2$ yields the Euclidean distance, $\mathbf{x}$ and $\mathbf{z}$ are two feature vectors each comprising p features, x_i and z_i are the values of the i^{th} feature for $\mathbf{x}$ and $\mathbf{z}$, p is the total number of features, w_x and w_z are weights assigned to the feature vectors $\mathbf{x}$ and $\mathbf{z}$. In SMOTE-N, the weights are removed from the equation. After identifying the k nearest neighbours, the feature values of the new synthetic minority class feature vector are determined by the most frequently occurring value among the k nearest neighbours. In this study, we used k = 5, as suggested by the developers of SMOTE (Chawla et al., 2002).

4.12. Imputation of missing values

Missing data are often met in clinical applications and are a major issue in medical research. Many machine learning algorithms including those described in the preceding sections cannot handle missing data and default to discarding missing values. Thus, observations which include missing data are usually eliminated and only complete cases (i.e., those with no missing data) are included in the analysis, a process known as *complete-case analysis*. Complete-case analysis may yield results which are biased considering the reasons why data are missing, and also reduces the original sample size which causes a reduction of precision and power. Missing data

mechanisms can be broadly categorised into three groups (Rubin, 1976): a) *Missing Completely At Random*: data are missing due to chance, that is missing values are independent of the observed or other missing values in the data. In this case, missing values can be ignored, and a complete-case analysis will not be biased, b) *Missing At Random*: the probability that a value of a particular variable X is missing is independent of the variable X itself but is related to the values of some other variable(s) in the study, and c) *Missing Not At Random*: the probability that a value of a random variable X is missing depends on that variable itself. In this case, a complete-case analysis would lead to biased results.

Instead of eliminating cases with missing data, missing values can be replaced with estimates of their true values based on other available information in a dataset, a process known as *imputation*. Following imputation, machine learning tools which are designed for complete data can be implemented. There are a number of imputation techniques, the description of which is beyond the scope of this thesis, but we refer interested readers to the book *"Handbook of missing data methodology"* (Molenberghs et al., 2014) for an extensive overview of missing data methods.

In this section, we will present a short summary of *multivariate imputation by chained equations* (Buuren & Groothuis-Oudshoorn, 2011) which we have used in this study, and which involves generating several plausible imputed versions of an original dataset which includes missing values. Multiple imputation by chained equations assumes that missing values are missing at random and entails the following steps:

1. Missing values for each feature in the dataset are filled in with temporary “place holders”, such as mean imputations (i.e., the missing values of a variable are replaced with the mean of the observed values for that variable).
2. For a given feature X, these temporary “place holders” are discarded. So, at this point, only feature X has missing values, while all other features have their “temporary place holders” in place of missing values.
3. A model is trained using the observed values of the feature X as the response, and the rest of the features as the predictor variables. Each feature is modelled according to its distribution, i.e., linear regression is used for continuous features, and logistic regression for categorical features.
4. The trained regression model is used to make predictions for the missing values of feature X. When feature X is subsequently used as a predictor for the rest of the features, then both its observed and its imputed values (by the regression model) will be used.
5. Steps 2 to 4 are repeated for all features with missing values, until all missing values have been imputed using regression models. The order that the features are usually considered for imputation is from the feature with the least missing values to the feature with the most missing values.
6. Steps 2 to 5 are repeated iteratively until convergence. The iterations required usually lie between 10 and 20.

The whole process (steps 1 to 6) is repeated h times (usually h is set between 5 and 10), so that in the end, h imputed datasets are generated which are identical regarding the observed values of the original dataset but differ in the imputed values (Azur et al., 2011). We need to note that multiple imputation by chained equations is applied using

only information from the training data, so that there is no data leakage. Subsequently, we perform data analysis and estimate the parameters of interest for each one of the imputed datasets. Finally, the estimated parameters are pooled into a single estimate.

4.13. Interpretation of the results of a machine learning model

After having developed a machine learning model, we are often interested in how the combination of the selected features influences the predicted outcome of the model. Graphs are useful interpretation tools, however, due to the limitations of computer graphics and human perception, they are limited to low-dimensional representations (i.e., it is difficult to plot in a single graph the predicted outcome as a function of all the features included in the model). Thus, a common solution for interpreting the results of a learner is to create a collection of graphs, each one illustrating the partial dependence of the outcome on a subset of the features used in the model. Thus, *partial dependence plots (PDP)* (Friedman, 2001) which show the marginal effect a small subset of features have on the outcome, are commonly used for interpretation of machine learning models. Let us consider a machine learning model comprising p features. Then $\mathbf{x}_s$ is a subvector, indexed by $s \subset \{1,2,\dots,p\}$, comprising the features (usually one or two) we are interested in (i.e., for which we need to define the impact on the predicted outcome of the model, that is a numerical value in regression settings or class probability in classification settings). Let $\mathbf{x}_c$ be a subvector, indexed by $c \subset \{1,2,\dots,p\}$, comprising the rest of the selected features in the model, such that $s \cup c = \{1,2,\dots,p\}$. Partial dependence functions work by marginalising the predicted outcome over the distribution of the feature vector $\mathbf{x}_c$, so that the function represents the impact of feature vector $\mathbf{x}_s$ on the predicted outcome. The partial function $\hat{f}_s$ is given by:

$$\hat{f}_s = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(\mathbf{x}_s, \mathbf{X}_{ic}\right) \tag{3.39}$$

where n is the number of observations in the training dataset, and $\mathbf{X}_{ic}$ represents the values of the features in $\mathbf{x}_c$ for the observations in the training dataset. The partial function shows what the average marginal effect on the predicted outcome of a machine learning model is, for given values of the features in $\mathbf{x}_s$. However, an assumption is that the features in $\mathbf{x}_s$ and $\mathbf{x}_c$ are not correlated.

4.14. Summary of the steps for data analysis

In this section, we will summarise the data analysis steps in the context of supervised learning problems in regression and classification settings, which we have described in the preceding sections. The methodology we have followed in this study comprises the following steps:

1. Begin by visualizing data in order to identify outliers, trends, and patterns. Draw scatterplots which allow for identification of potential relationships between two variables.
2. Compute correlation coefficients in order to quantify the strength of the association between pairs of variables.
3. Rank the features of a given dataset from the most important to the least important one using feature selection techniques.
4. Apply parametric statistical learning tools (e.g., linear regression for continuous responses or logistic regression for classification tasks) which set a useful

benchmark which we subsequently explore whether we can improve upon by applying non-parametric machine learning algorithms, such as RF.

5. Calculate the performance of the machine learning model as a function of the number of features used and keep the statistical model which has the greatest predictive power with the fewest possible features. Comparison between models can be achieved using statistical hypothesis tests.
6. In the context of imbalanced classification, use data balancing techniques, such as SMOTE, in order to mitigate the class imbalance problem and use appropriate performance metrics.
7. If data are missing at random or completely at random, impute missing values using, for example, multiple imputation by chained equations.
8. Validate the model using an additional dataset, or, if not available, using *k*-fold cross-validation.
9. Surrogate testing to test whether the selected model significantly outperforms a naïve benchmark.
10. Interpret the results.

The following chapters use the methodology outlined above to address the objectives of the study: identifying the early life risk factors that predict global and regional brain volumes at term-equivalent age (Chapter 5), as well as language outcomes at two years CGA following preterm birth (Chapter 6).

CHAPTER 5: PREDICTION OF GLOBAL AND REGIONAL CEREBRAL VOLUMES FOLLOWING PRETERM BIRTH

This chapter presents a study we undertook to address the first objective of the current thesis: to identify the early life risk factors that impact brain growth following preterm birth by developing machine learning models that accurately predict global and regional cerebral tissue volumes at term–equivalent age.

5.1. Introduction

The foetal brain undergoes rapid growth during the third trimester of gestation and is particularly vulnerable to insults, which can lead to aberrant development (see section 2.1.1). Groupwise studies using volumetric MRI report that preterm birth is associated with impaired total, white matter, cortical grey matter, deep nuclear grey matter, and cerebellar growth (Ball et al., 2012; Batalle et al., 2018; Boardman et al., 2006; Inder et al., 2005; Limperopoulos et al., 2005; Loh et al., 2017; Mewes et al., 2006; Srinivasan et al., 2007; Thompson et al., 2007), although growth failure is not inevitable (Boardman et al., 2007) which suggests there is considerable individual variation. In turn, altered brain volume following preterm birth has been associated with various short- and long-term neurodevelopmental deficits (for details, see section 2.1.3). Understanding the risk factors associated with altered brain growth could lead to more timely identification of children who are at high risk of neurodevelopmental impairment and who may benefit from early intervention programmes.

The objective of the current study was to identify the combination of early life risk factors that predict total brain volume, and regional brain tissue volumes (white matter, cortical grey matter, deep nuclear grey matter, and cerebellum). Different brain regions are characterised by different growth rates (Nishida et al., 2006), and thus, we hypothesised that they are differentially vulnerable to perinatal exposures. Previous investigations of variables that predict brain volumes at term–equivalent age following preterm birth have used linear regression (Alexander et al., 2019; Ball et al., 2012; Belfort et al., 2016; Brouwer, Kersbergen, van Kooij, et al., 2017; Granger et al., 2018; Guillot et al., 2020; Inder et al., 2005; Kidokoro et al., 2014; Limperopoulos et al., 2005; Matthews et al., 2018; Nguyen The Tich et al., 2011; Pecheva et al., 2019; Power et al., 2019; Sveinsdóttir et al., 2018; Thompson et al., 2008; Thompson et al., 2019b), linear mixed effects models (Kamino et al., 2019; Knickmeyer et al., 2017; Thompson et al., 2007), generalised linear models (Boardman et al., 2007), generalised least squares models (Zwicker et al., 2016), general linear models (Duerden et al., 2020), generalised estimating equation (J. Schneider et al., 2018), logistic regression (Inder et al., 2003), Spearman rank correlation (Dimitrova et al., 2021), and canonical correlation analysis (Ball et al., 2017). However, often the underlying relationship between the predictors and the response is more complicated, such that linear models and correlation analysis may not provide an accurate representation of the underlying algorithmic relationship. Applying machine learning techniques to a cohort of preterm infants that is phenotyped with brain imaging, clinical, and socioeconomic information enables the exploration of non-linear relationships between the response and the predictor variables. Early identification of the combination of clinical, demographic, and environmental perinatal variables that accurately predicts brain tissue volumes at

term–equivalent age could potentially lead to the improvement of neuroprotective strategies and perinatal treatments.

5.2. Materials and methods

5.2.1. Participants

Participants were preterm infants born before 33 completed weeks of gestation at the Royal Infirmary of Edinburgh between October 2016 and September 2021 and recruited as part of the TEBC (Boardman et al., 2020) (see Section 3.3). The inclusion and exclusion criteria for the study are described in detail in Section 3.3.

5.2.2. Clinical and sociodemographic features

The selection of clinical and demographic features used in subsequent analysis was guided by extant literature linking early life risk factors with altered brain growth in preterm infants (see section 2.1.2). Specifically, we studied the contribution towards prediction of cerebral tissue volumes of the following features: GA at birth (based on first trimester ultrasound) (Anjari et al., 2009; Ball et al., 2010, 2012; Boardman et al., 2006; Inder et al., 2005; Kidokoro et al., 2014; Limperopoulos et al., 2005; Partridge et al., 2004), GA at MRI, birthweight z-score (a measure of birthweight standardised for age and sex based on the INTERGROWTH-21st international standards (Villar et al., 2014, 2016)) (Alexander et al., 2019; Knickmeyer et al., 2016; Matthews et al., 2018; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Thompson et al., 2019b), sex (Alexander et al., 2019; Dibble et al., 2021; Gilmore et al., 2007; Kersbergen et

al., 2016; Matthews et al., 2018; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Ruigrok et al., 2014; Thompson et al., 2007; Thompson et al., 2019b), multiple birth (Alexander et al., 2019; Thompson et al., 2019a), early or late-onset sepsis (defined as blood stream infection with (a) bacterial pathogen isolated from blood culture or (b) blood culture growing coagulase-negative staphylococcus, along with one or more signs of generalised infection, and treatment with intravenous antibiotics for ≥5 days; early-onset is within 72 hours of birth and late-onset is after 72 hours) (Matthews et al., 2018), NEC (requiring medical [seven days nil by mouth] or surgical management) (Kidokoro et al., 2014; Matthews et al., 2018; Pogribna et al., 2013; Shah et al., 2008), BPD (defined as oxygen requirement at ≥36 weeks CGA) (Anjari et al., 2009; Ball et al., 2010; Boardman & Counsell, 2020; Inder et al., 2005; Kidokoro et al., 2014; Thompson et al., 2007),duration of intubation while in the NICU (Boardman et al., 2007; Brouwer, Kersbergen, Van Kooij, et al., 2017; Guillot et al., 2020; Nguyen The Tich et al., 2011; Pogribna et al., 2013; Rogers et al., 2016), duration of parenteral nutrition (Brouwer, Kersbergen, van Kooij, et al., 2017; Kidokoro et al., 2014), breast milk exposure (defined as the proportion of in-patient days receiving exclusive maternal and/or donor breast milk) (Belfort et al., 2016; Blesa et al., 2019; Sullivan et al., 2022), smoking during pregnancy (at any point during pregnancy, self-reported by parents after birth) (Ekblad et al., 2015), and socioeconomic status of the family (Thompson et al., 2019b) operationalised as Scottish Index of Multiple Deprivation 2016 (SIMD2016) quintile which comprises five categories (1-5), where 1 indicates the most deprived and 5 indicates the least deprived (Scottish National Statistics, 2016). $MgSO_4$ is recommended in clinical practice guidelines worldwide for women at risk of very preterm birth for fetal neuroprotection (World Health Organization, 2015) though the mechanism of its neuroprotective effects is not well understood. Thus, we

investigated the effect of exposure to $MgSO_4$ on brain volumes at term–equivalent age. In addition, we included maternal depression and anxiety (data was obtained from maternal pregnancy records: previous history and/or during current pregnancy), as psychological distress has been associated with atypical brain development (Dean et al., 2018; Gentile, 2017; Rifkin-Graboi et al., 2015) and a range of adverse neurodevelopmental outcomes in offspring (Boardman & Counsell, 2020; Davis & Sandman, 2012; Grizenko et al., 2012; Robinson et al., 2011). Finally, we aimed to assess the effect of exposure to antenatal corticosteroids (any antenatal corticosteroids or a complete course of antenatal corticosteroids [defined as two doses 24h apart]) on cerebral tissue volumes, because antenatal corticosteroid administration for acceleration of fetal lung maturation has been associated with lower risk of neurodevelopmental impairment following preterm birth (S. Chawla et al., 2016; McGoldrick et al., 2020; Tyson et al., 2008; Valavani et al., 2021; Vesoulis et al., 2018).

5.2.3. Image acquisition and analysis

Infants underwent a brain MRI scan at term–equivalent age. Details on the acquisition of sMRI are provided in Section 3.6. T2-weighted MRI scans were acquired for all subjects and neonatal total brain, cortical and deep grey matter, white matter, and cerebellar volumes were obtained using the minimal processing pipeline of the developing human connectome project (dHCP) (Makropoulos et al., 2014, 2018).

5.3. Data analysis

5.3.1. Statistical mapping

We used the RF algorithm, operating in regression mode for the estimation of the cerebral tissue volumes at term–equivalent age following preterm birth in the study, using the standard default parameters (500 trees, optimizing the Gini index, and searching for the best split at each tree node from the randomly selected features equal to the square root of the original number of features) (for more details on RF, see Section 4.7.3.3). We compared two feature selection algorithms, which are optimised and tuned for the RF learner: (a) RF variable importance (Hastie et al., 2009), and (b) Boruta (Kursa & Rudnicki, 2010), which rank the features of a given dataset based on their contribution towards prediction of the response variable (see Sections 4.5.2 and 4.5.3). The final feature ranking for each feature selection algorithm was determined using LOOCV, using only the training data set in each cross-validation iteration and following the process described in Section 4.6. Ultimately, we obtained a vector which consisted of feature indices ordered from the most important to the least important one, for each feature selection algorithm. These features were then used to train the RF learner (Breiman, 2001) using a progressively increasing number of features (i.e. only the first selected features, the top two selected features, the top three, etc.), so that the performance of the model was calculated as a function of the number of features used. Following the principle of parsimony, we selected the most parsimonious model, i.e. the statistical model which had the greatest predictive power with the fewest possible features.

5.3.2. Model performance

Model validation was implemented using LOOCV. Missing data for both numeric and categorical features were imputed using multiple imputation by chained equations (Raghunathan et al., 2001; Van Buuren, 2007), based only on the information in the training set, independently in each LOOCV iteration (see Section 4.11). The performance of the model was evaluated using the RMSE. We also report the MAE and MAPE of the models, as additional, intuitively understandable performance measures (for more details on these metrics, refer to Section 4.8.1).

5.3.3. Surrogate testing and interpretation of findings

Subsequently, the RMSE of the selected model was compared to the RMSE of the naïve (surrogate) model (the predicted response is equal to the mean value of the response) using a paired t-test (see Section 4.9.2.1). The assumption of the normal distribution of the differences between the pairs was tested using a normal probability quantile-quantile plot. In case this assumption was violated, the non-parametric alternative Wilcoxon signed-rank test was performed. Statistical significance was set at $\alpha=0.05$. This is useful to demonstrate whether the machine learning model performs over and above a naïve model. PDP (Friedman, 2001) were constructed to assess how the selected features influence the predicted outcome of the RF learner (see Section 4.12). Data analysis was conducted in R.

5.4. Results

In total, 169 preterm (median GA ± Interquartile Range [IQR]: 30 ± 3.43 weeks) underwent a brain MRI at term–equivalent age. The characteristics of the cohort are presented in Table 5.1. The overall percentage of missing values in the dataset was 0.3% (data were missing on BPD and breast milk exposure for one infant, on duration of non-enteral feeds, breast milk exposure, and smoking during pregnancy for one infant, on BPD, duration of non-enteral feeds, and breast milk exposure for one infant, and two infants were missing data on maternal smoking during pregnancy).

Table 5.1. Characteristics of the cohort.

Characteristics	**Preterm infants (N=169)**
GA at birth (weeks)	30±3.43 (22.14–32.86)
GA at MRI scan (weeks)	40.71±1.86 (36.57–45.86)
Birth weight (grams)	1300±540 (370–2510)
Birth weight z-score	0.13±1.05 (-3.13–2.14)
Sex Male Female	 95(56) 74(44)
$MgSO_4$	134(79)
Antenatal corticosteroids (any)	161(95)
Antenatal corticosteroids (complete course)	113(67)
Multiple birth	50(30)
Bronchopulmonary dysplasia	41(25)
Early onset sepsis	13(8)
Late onset sepsis	27(16)
Necrotizing enterocolitis	7(4)
Retinopathy of prematurity	7(4)
Duration of intubation (days)	0±2 (0–63)
Duration of exclusive breast milk exposure (days)	32.5±30 (0–122)
Duration parenteral nutrition (days)	7±4 (0-65)
Maternal age (years)	32±8

	(17–45)
Maternal smoking during pregnancy	30(18)
Maternal depression	33(20)
Maternal anxiety	41(24)
SIMD2016 quintile	
1	32(19)
2	32(19)
3	28(16)
4	33(20)
5	44(26)

Variables are presented in the form of median ± IQR (range) or number (%).

5.4.1. Prediction of total brain volume in preterm infants

Figure 5.1.A illustrates the out-of-sample performance of the RF regression model for the prediction of total brain volume as a function of the number of features selected by the RF variable importance and Boruta algorithms. The best performing feature size (minimum RMSE) is six using Boruta (RMSE=29.23cm^3, MAE=23.89cm^3, MAPE=6.57%). Comparing the selected model to the naïve model, we found that the former is significantly better than the latter (d=0.46, p<0.0001). The selected feature subset comprises birthweight z-score, GA at birth, GA at MRI scan, sex, duration of intubation, and SIMD2016 quintile which are the jointly most predictive features towards the prediction of total brain volume in preterm infants. The PDP show that total brain volume increases with increasing birthweight z-score (Fig. 5.1.B), GA at birth and MRI scan (Fig. 5.1.C – D). Female sex, longer duration of intubation, and lower socioeconomic status were associated with lower total brain volume (Fig. 5.1.E – G).

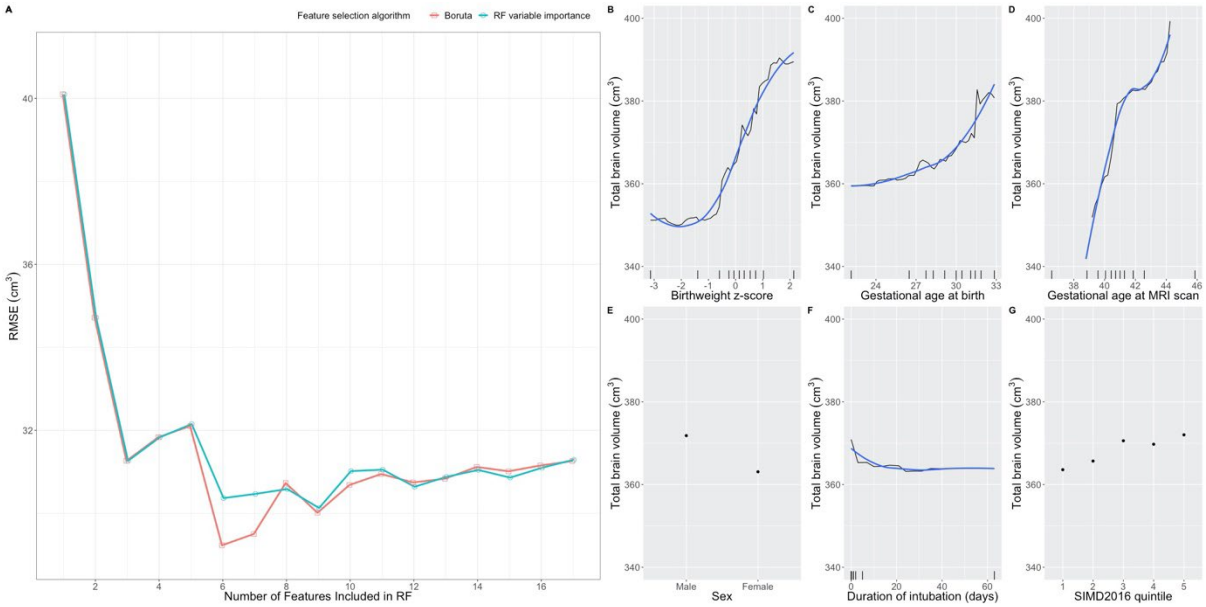


Figure 5.1. Selected model and PDP for total brain volume. (A) Comparison of out-of-sample LOOCV RMSE results of the RF prediction model of total brain volume using the features selected by the Boruta and RF variable importance algorithms; (B) PDP for the effect of birthweight z-score on total brain volume; (C) PDP for the effect of gestational age at birth on total brain volume; (D) PDP for the effect of gestational age at MRI scan on total brain volume; (E) PDP for the effect of sex on total brain volume; (F) PDP for the effect of duration of intubation on total brain volume; (G) PDP for the effect of SIMD2016 quintile on total brain volume.

5.4.2. Prediction of white matter volume in preterm infants

Figure 5.2.A illustrates the out-of-sample performance of the RF regression model for the prediction of white matter volume as a function of the number of features selected by the RF variable importance and Boruta algorithms. The best performing feature size (minimum RMSE) is seven using RF variable importance, but this is not a statistically significant improvement over the RMSE when using only four features selected by the RF variable importance or Boruta algorithm (p=0.2105). Hence, following the principle of parsimony, we choose the least number of features giving the most accurate results according to the RMSE. We choose to keep the RF model with four features (RMSE=14.92cm^3, MAE=11.98cm^3, MAPE=7.16%). Comparing the selected model to the naïve model, we found that the former is significantly better than the latter (d=0.30, p<0.0001). The selected feature subset comprises birthweight z-score, GA at birth, GA

at MRI scan, and sex. The PDP show that white matter volume increases with increasing birthweight z-score (Fig. 5.2.B), GA at birth and MRI scan (Fig. 5.2.C – D). Moreover, male sex was associated with greater white matter volume (Fig. 5.2.E).

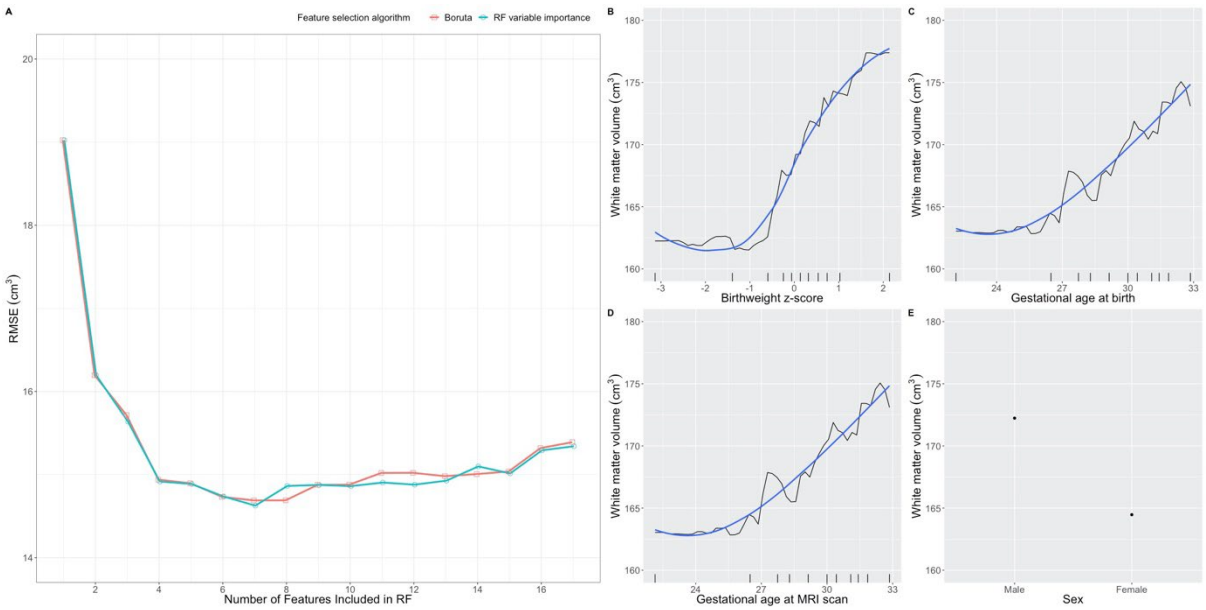


Figure 5.2. Selected model and PDP for white matter volume. (A) Comparison of out-of-sample LOOCV RMSE results of the RF prediction model of white matter volume using the features selected by the Boruta and RF variable importance algorithms; (B) PDP for the effect of birthweight z-score on white matter volume; (C) PDP for the effect of gestational age at birth on white matter volume; (D) PDP for the effect of gestational age at MRI scan on white matter volume; (E) PDP for the effect of sex on white matter volume.

5.4.3. Prediction of cortical grey matter volume in preterm infants

Figure 5.3.A illustrates the out-of-sample performance of the RF regression model for the prediction of cortical grey matter as a function of the number of features selected by the Boruta and RF variable importance algorithms. We selected the model comprising three features using the Boruta or RF variable importance algorithm (RMSE=14.43cm^3, MAE=11.45cm^3, MAPE=8.47%), which was significantly better than the naïve model (d=0.47, p<0.0001). The selected feature subset comprises birthweight z-score, GA at birth, and GA at MRI scan which are the jointly most

predictive features towards the prediction of cortical grey matter volume in preterm infants. The PDP show that cortical grey matter volume increases with increasing birthweight z-score (Fig. 5.3.B) and GA at birth and MRI scan (Fig. 5.3.C – D).

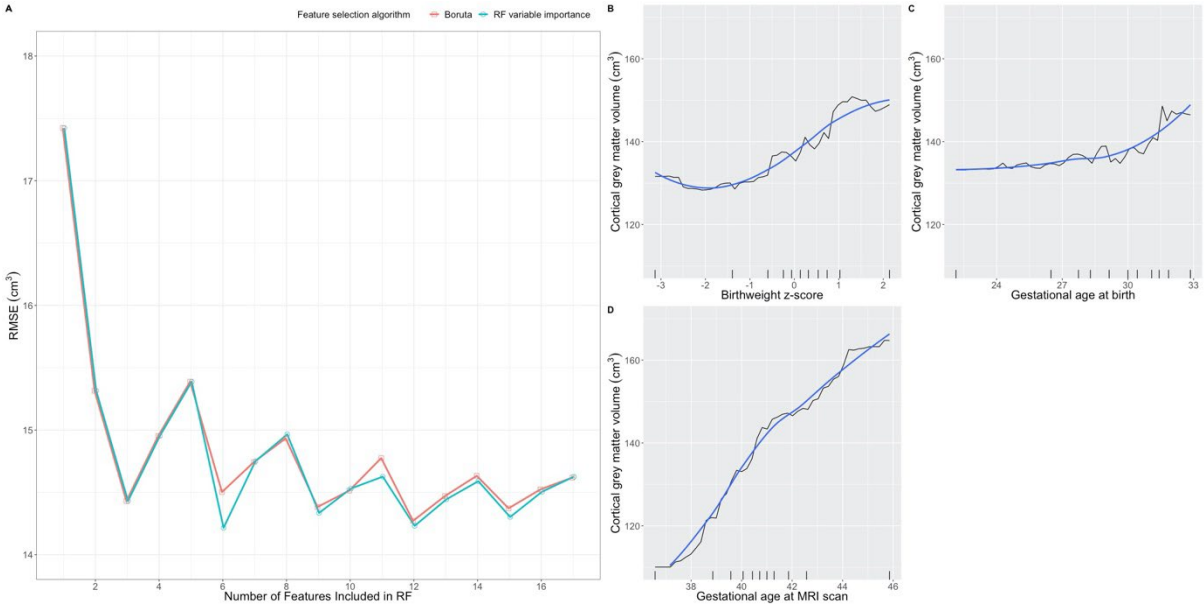


Figure 5.3. Selected model and PDP for cortical grey matter volume. (A) Comparison of out-of-sample LOOCV RMSE results of the RF prediction model of cortical grey matter volume using the features selected by the Boruta and RF variable importance algorithms; (B) PDP for the effect of birthweight z-score on cortical grey matter volume; (C) PDP for the effect of gestational age at birth on cortical grey matter volume; (D) PDP for the effect of gestational age at MRI scan on cortical grey matter volume.

5.4.4. Prediction of deep grey matter volume in preterm infants

Figure 5.4.A illustrates the out-of-sample performance of the RF regression model for the prediction of deep grey matter volume as a function of the number of features selected by the Boruta and RF variable importance algorithms. Following the principle of parsimony, we chose to keep the RF model with six features using the RF variable importance algorithm (RMSE=1.92cm^3, MAE=1.45cm^3, MAPE=5.54%). Comparing the selected model to the naïve model, we found that the former is significantly better than the latter (d=0.33, p<0.0001). The selected feature subset comprises birthweight

z-score, GA at birth, GA at MRI scan, duration of intubation, sex, and breast milk exposure which are the jointly most predictive features towards the prediction of deep grey matter volume in preterm infants. The PDP show that deep grey matter volume increases with increasing birthweight z-score (Fig. 5.4.B), GA at birth and MRI scan (Fig. 5.4.C – D), and higher breast milk exposure (Fig. 5.4.G). Prolonged duration of intubation and female sex are associated with lower deep grey matter volume (Fig. 5.4.E – F).

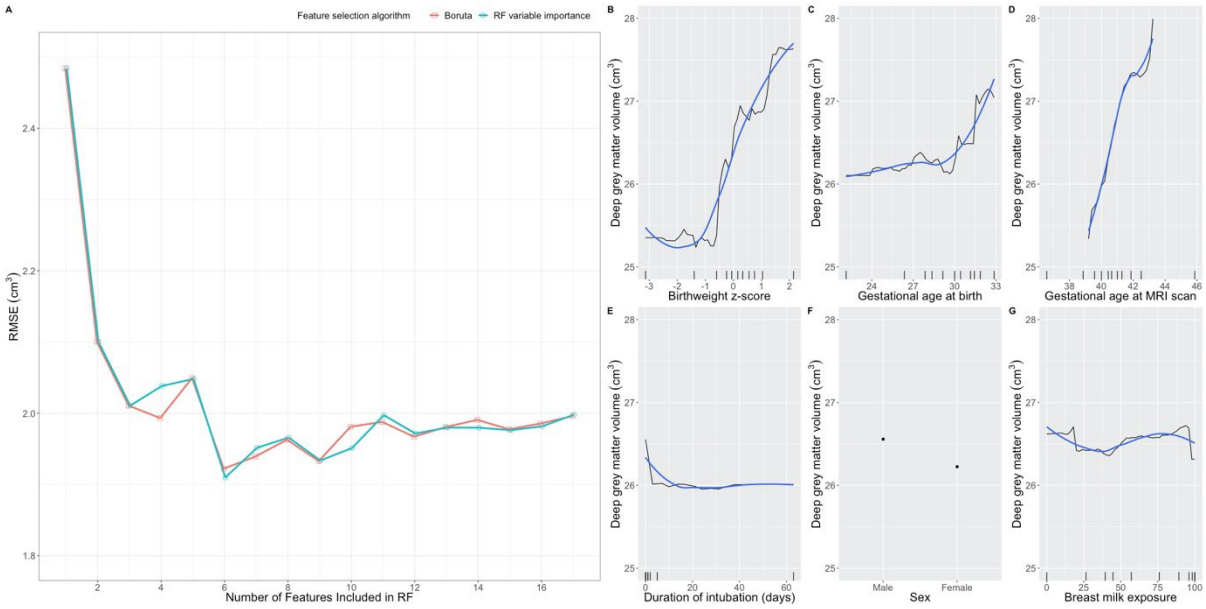


Figure 5.4. Selected model and PDP for deep grey matter volume. (A) Comparison of out-of-sample LOOCV RMSE results of the RF prediction model of deep grey matter volume using the features selected by the Boruta and RF variable importance algorithms; (B) PDP for the effect of birthweight z-score on deep grey matter volume; (C) PDP for the effect of gestational age at birth on deep grey matter volume; (D) PDP for the effect of gestational age at MRI scan on deep grey matter volume; (E) PDP for the effect of duration of intubation on deep grey matter volume; (F) PDP for the effect of sex on deep grey matter volume; (G) PDP for the effect of breast milk exposure on deep grey matter volume.

5.4.5. Prediction of cerebellar volume in preterm infants

Figure 5.5.A illustrates the out-of-sample performance of the RF regression model for the prediction of cerebellar volume as a function of the number of features selected by the different feature selection algorithms. These data show that feeding a subset of six features selected by the RF variable importance algorithm to the RF regression model gives the lowest RMSE (RMSE=3.08cm^3, MAE=2.18cm^3, MAPE=8.75%). Comparing the selected model to the naïve model, however, we found that the former is significantly better than the latter (d=0.22, p=0.0142). The selected feature subset comprises birthweight z-score, GA at birth, GA at MRI scan, sex, SIMD2016 quintile and duration of parenteral nutrition which are the jointly most predictive features towards the prediction of cerebellar volume in preterm infants. The PDP show that cerebellar volume increases with increasing birthweight z-score (Fig. 5.5.B), GA at birth and MRI scan (Fig. 5.5.C – D), and higher socioeconomic status (Fig. 5.5.E). In addition, longer duration of parenteral nutrition and female sex are associated with lower cerebellar volume (Fig. 5.5.F – G).

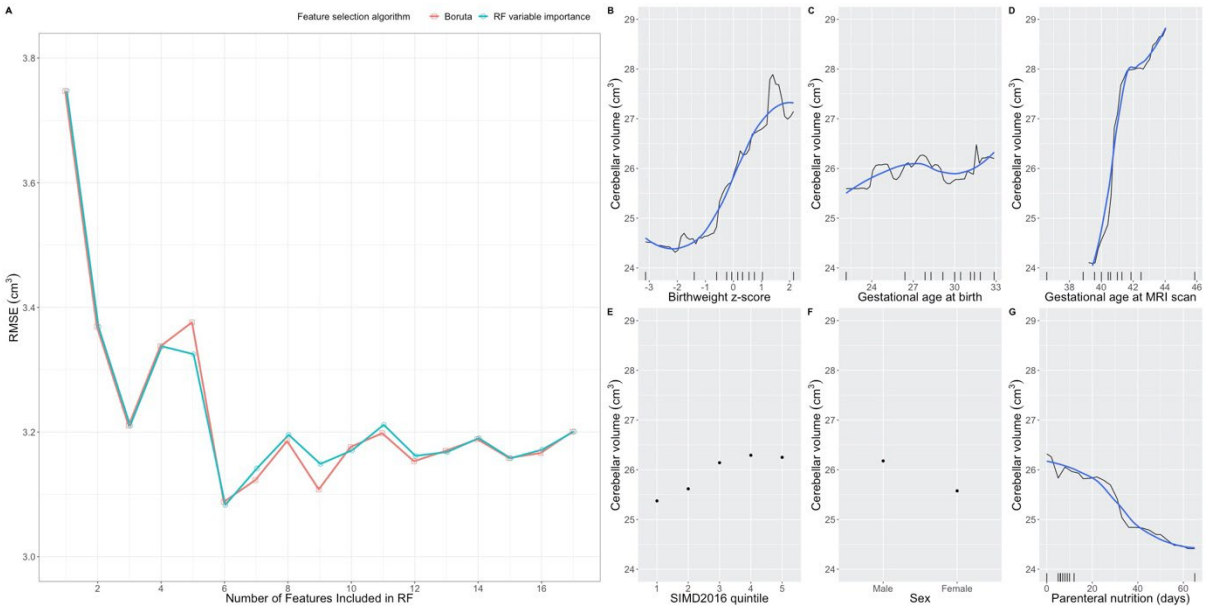


Figure 5.5. Selected model for cerebellar volume. (A) Comparison of out-of-sample LOOCV RMSE results of the RF prediction model of cerebellar volume using the features selected by the Boruta and RF variable importance algorithms; (B) PDP for the effect of birthweight z-score on cerebellar volume; (C) PDP for the effect of gestational age at birth on cerebellar volume; (D) PDP for the effect of gestational age at MRI scan on cerebellar volume; (E) PDP for the effect of SIMD2016 on cerebellar volume; (F) PDP for the effect of sex on cerebellar volume; (G) PDP for the effect of duration of parenteral nutrition on cerebellar volume.

5.4.6. Summarizing findings

The findings on the selected feature subsets and performance measures for cerebral tissue volumes following preterm birth are summarised in Table 5.2. Figure 5.6 presents the importance attributed to the selected features by the feature selection algorithm used in each model.

Table 5.2. Selected feature subsets and performance measures for prediction of cerebral tissue volumes following preterm birth. The features are presented in rank order from the most important to the least important.

Cerebral volumes	Selected features	RMSE (cm^3)	MAE (cm^3)	MAPE (%)
Total brain	1. GA at MRI scan 2. Birthweight z-score 3. GA at birth 4. Sex 5. Duration of intubation 6. SIMD2016 quintile	29.23	23.89	6.57
White matter	1. Birthweight z-score 2. Sex 3. GA at birth 4. GA at MRI scan	14.92	11.98	7.16
Cortical grey matter	1. GA at MRI scan 2. Birthweight z-score 3. GA at birth	14.43	11.45	8.47
Deep grey matter	1. GA at MRI scan 2. Birthweight z-score 3. GA at birth 4. Duration of intubation 5. Sex 6. Breast milk exposure	1.92	1.45	5.54
Cerebellum	1. GA at MRI scan 2. Birthweight z-score 3. SIMD2016 quintile 4. Sex 5. Duration of parenteral nutrition 6. GA at birth	3.08	2.18	8.75

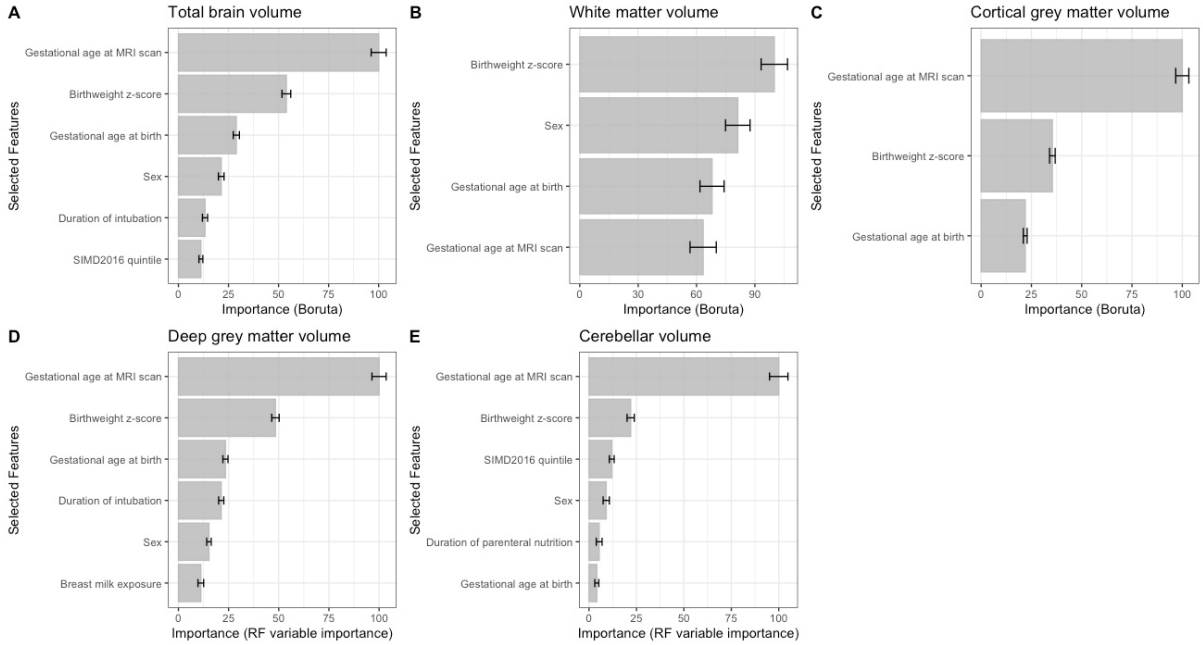


Figure 5.6. Feature importance plots with error bars. Importance attributed to each feature of the selected feature subsets for prediction of A) total brain volume, B) white matter volume, C) cortical grey matter volume, D) deep grey matter volume, and E) cerebellar volumes. Feature importance is expressed relative to the maximum.

5.5. Discussion

We used advanced machine learning techniques to identify the perinatal variables that are jointly associated with brain volume in a cohort of preterm infants. The results of the study show that birthweight z-score, GA at birth, age at scan, and sex, in combination with postnatal nutrition, respiratory morbidity, and socioeconomic status affect cerebral tissue volumes at term–equivalent age following preterm birth. In line with our initial hypothesis, we have demonstrated that different combinations of clinical and environmental exposures influence the morphology of brain tissue classes.

Birthweight z-score, GA at birth, and age at MRI scan influence the volume of all brain regions under study following preterm birth. Higher birthweight z-score, which reflects better fetal growth, was associated with greater volumes in all regions. This is

consistent with earlier findings of strong positive associations between birthweight z-score and global brain volumes at term–equivalent age (Alexander et al., 2019; Knickmeyer et al., 2017; Matthews et al., 2018; Nguyen The Tich et al., 2011; Thompson et al., 2019a). GA at birth was also an important predictor of total brain volume, white matter, cortical grey matter, deep nuclear grey matter, and cerebellar volumes; older GA was associated with larger brain volumes. This pattern of greater volume reduction with an increasing degree of prematurity aligns with the findings of previous reports which have also shown positive associations between GA at birth and regional brain volumes at term–equivalent age (Ball et al., 2012; Boardman et al., 2006; Inder et al., 2005). Other studies, however, have found no associations between GA and total tissue volume following preterm birth, which suggests there is considerable individual variation (Alexander et al., 2019; Boardman et al., 2007).

Sex was an important predictor of overall brain volume, white matter, deep nuclear grey matter, and cerebellar volumes at term–equivalent age, with male sex being associated with greater volumes. Our findings replicate previous studies showing that sexual dimorphism is present at birth, with males having larger global brain volumes than females (Alexander et al., 2019; Gilmore et al., 2007; Kersbergen et al., 2016; Matthews et al., 2018; Nguyen The Tich et al., 2011; Ruigrok et al., 2014; Thompson et al., 2007, 2019b), and that these differences are sustained through life (Reiss et al., 2004; Ruigrok et al., 2014). Sexual dimorphism of the brain may be mediated by the effects of gonadal hormones, glucocorticoids, and genetic mechanisms on brain development during the perinatal period (Le Dieu-Lugon et al., 2020; McCarthy, 2008; Owen & Matthews, 2003; Stoye et al., 2020).

Our findings demonstrate that a longer duration of intubation while in the NICU is associated with reduced total brain and deep grey matter volume. The duration of intubation is a proxy for the severity of respiratory illness, so the results provide further evidence for an association between lung disease and brain development. This finding aligns with previous studies which have also shown the negative impact of respiratory illness on brain growth at term and that prolonged duration of assisted ventilation and oxygen exposure adversely affects cerebral development leading to decreased cerebral volumes (Boardman et al., 2007; Kersbergen et al., 2016; Nguyen The Tich et al., 2011; Thompson et al., 2007). Ball et al. (2010) have demonstrated that BPD also affects the microstructure of the brain. These results together indicate that respiratory illness is an important risk factor for adverse neurocognitive development in preterm infants.

This study highlights the important role of early postnatal nutrition for improved brain development after preterm birth. In this cohort of preterm infants, prolonged duration of parenteral nutrition was associated with smaller cerebellar volume, in line with previous reports in the literature of negative association between the duration of non–enteral feeds and cerebral tissue volumes at term–equivalent age (Binder et al., 2021; Brouwer, Kersbergen, van Kooij, et al., 2017; Coviello et al., 2018; Kidokoro et al., 2014; Parikh et al., 2013). The cerebellum is particularly vulnerable to environmental exposures, including nutrition, thought to be due to its exponential growth during the third trimester of gestation (Volpe, 2009). Optimal nutrient intake during the neonatal period is crucial for enhanced brain growth and maturation (Beauport et al., 2017; Boardman & Counsell, 2020; Coviello et al., 2018; J. Schneider et al., 2018). However, preterm infants requiring prolonged parenteral nutrition, which usually reflects a more

severe neonatal course, are at increased risk of nutritional deficits (Binder et al., 2018; Wang et al., 2021), which can lead to altered brain structure (Ramel & Georgieff, 2014).

We found that higher exposure to breast milk feeding while in the NICU is positively associated with deep nuclear grey matter volume. Studies in the extant literature have shown that breast milk feeding following preterm birth, as opposed to formula feeds, is associated with improved white matter and cortical maturation (Blesa et al., 2019; Pogribna et al., 2013; Schneider et al., 2018; Sullivan et al., 2022), and improved neurodevelopmental outcomes at pre-school age and beyond (Belfort et al., 2016; Lechner & Vohr, 2017; Luby et al., 2016; J. Miller et al., 2018; Parker et al., 2021). Our study supports recent work demonstrating that a longer duration of breast milk feeding in early life is associated with greater regional brain volumes, especially in the deep nuclear grey matter, amygdala, hippocampus, and cerebellum (Belfort et al., 2016; Belfort & Inder, 2022; Ottolini et al., 2020), offering a potential mechanism for the beneficial effects of breast milk feeding on neurodevelopmental outcomes. Luby et al. (Luby et al., 2016) have previously reported that the beneficial effects of breastmilk feeding on preschool children's Intelligence Quotient (IQ) are mediated through subcortical grey matter volume. The results are consistent with a growing body of literature showing that breast milk nutrition is associated with a favourable pattern of brain development in preterm infants.

Finally, the results show that neighbourhood deprivation is associated with decreased total brain volume at term–equivalent age, which is driven by a reduction of cerebellar volume. Most studies investigating the relationship between social factors and cerebral

tissue volumes have focused on school-aged children and adolescents (Hanson et al., 2011, 2013; Jednoróg et al., 2012; Merz et al., 2018). Our study, however, is one of the very few studies to have explored the association between socioeconomic status and brain structure at term–equivalent age (Betancourt et al., 2016; Jha et al., 2019; Knickmeyer et al., 2017; Lu et al., 2021; Thompson et al., 2019b; Triplett et al., 2022). Previous studies have demonstrated that the prefrontal cortex, hippocampi, and amygdalae are the most susceptible brain regions to the influence of socioeconomic disadvantage, due to their high levels of glucocorticoid receptors (S. B. Johnson et al., 2016). However, a recent study by Lu et al. (Lu et al., 2021) investigating fetal brain development, as well as studies on preschool and school-age children, and healthy adults have shown that early deprivation affects the development of the cerebellum (Bauer et al., 2009; Cavanagh et al., 2013; Stiver et al., 2015). The pathway remains unclear, but high levels of stress during pregnancy (S. B. Johnson et al., 2016), poor nutrition (Cortés-Albornoz et al., 2021), limited access to proper prenatal care and health services (Lee et al., 2016; Paredes et al., 2005), as well as exposure to environmental toxins or harmful substances such as air pollution (Herting et al., 2019) and cigarette smoking during pregnancy (Ekblad et al., 2015) are possible mechanisms explaining the impact of neighbourhood deprivation on total brain volume and cerebellar volume during late gestation and the early postnatal period.

This study is the first to develop advanced machine learning models for the prediction of total brain, white matter, cortical and deep nuclear grey matter, and cerebellar volumes. The main strength of our study is that we had a longitudinal cohort of preterm infants that enabled us to investigate a large number of clinical, demographic, and social variables. In addition, the use of non-parametric techniques enabled us to

explore non-linear relationships between brain volumes and the predictor variables, and thus, to accurately identify the combination of pre- and perinatal exposures that affect brain growth. We acknowledge some limitations in our study. The sample size is relatively small, and some features (i.e., NEC, ROP) may have not been selected as important due to low frequency in our sample. Finally, we have not been able to test replication in an independent cohort due to differences in data harmonisation inherent to multisite scanning.

This study revealed that a combination of clinical and environmental exposures best predicts cerebral tissue volumes at term–equivalent age following preterm birth. Potentially modifiable risk factors, including postnatal nutrition, respiratory morbidity, and socioeconomic status of the family, affect the brain volume of preterm infants at term–equivalent age. Neuroprotective strategies and preventive interventions should aim to optimise nutritional support of preterm infants, use ventilation strategies that minimise respiratory morbidity, and alleviate parental socioeconomic hardships.

In addition to perinatal practices, which significantly impact brain development in preterm infants, long-term neurodevelopmental outcomes can also be improved through developmental programmes and support services during early childhood (Spittle et al., 2015). So, our next aim was to develop a machine learning model comprising perinatal clinical, environmental, and brain imaging features to timely identify preterm infants at high risk of language impairment, who may benefit from targeted early interventions.

CHAPTER 6: PREDICTION OF LANGUAGE OUTCOME FOLLOWING PRETERM BIRTH

6.1. Introduction

This chapter comprises an article titled “Language Function Following Preterm Birth: prediction using Machine Learning”, published in the peer-reviewed journal *Pediatric Research.* The findings of this study were also presented at the Pediatric Academic Societies (PAS) 2021 Virtual Meeting, where it received the Student Research Award from the Society of Pediatric Research.

I, Evdoxia Valavani, conceived the idea of the published article with the help of co-authors, Athanasios Tsanas and James P Boardman. Additionally, I wrote the manuscript, analyzed the data and interpreted the results. Manuel Blesa and Paola Galdi conducted the image analysis. Co-authors, Manuel Blesa, Paola Galdi, Bethan Dean, Hilary Cruicksank, and Magdalena Sitko-Rudnicka collected the data. All authors edited the manuscript and approved the final version of the published article.

The current study has contributed to address the overarching objective of this thesis, which is to identify the early life exposures that affect the development of the preterm brain and subsequent neurodevelopmental outcomes. Specifically, our aim has been to identify which perinatal factors are predictive of future language development, thus timely detecting which preterm infants are at high risk of delay and potentially improving perinatal practices and targeting interventions.

This study addresses the second objective of the thesis, which is to develop a machine learning model that accurately predicts which preterm infants are at high risk of language impairment at two years CGA and who may benefit from targeted early interventions and support services. We tested our initial hypothesis that a machine learning model combining clinical, environmental, and brain imaging features would outperform existing statistical models in predicting language outcomes following preterm birth (see Section 2.4). This chapter presents in detail the methodology used, the findings, as well as the clinical and research implications of the study.

6.2. Published journal manuscript

Language Function Following Preterm Birth:

prediction using Machine Learning

Evdoxia Valavani[1], Manuel Blesa[2], Paola Galdi[2], Gemma Sullivan[2], Bethan Dean[2], Hilary Cruickshank[3], Magdalena Sitko-Rudnicka[4], Mark E. Bastin[5], Richard F. M. Chin[6,7], Donald J. MacIntyre[8], Sue Fletcher-Watson[9], James P. Boardman[2,5], and Athanasios Tsanas[1]

[1] Usher Institute, Medical School, University of Edinburgh, Edinburgh, UK

[2] MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, UK

[3] NHS Lothian-Neonatal Physiotherapy, Royal Infirmary of Edinburgh, Edinburgh, UK

[4] NHS Lothian-Neonatology, Royal Infirmary of Edinburgh, Edinburgh, UK

[5] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[6] Muir Maxwell Epilepsy Centre, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[7] Royal Hospital for Sick Children, Edinburgh, UK

[8] Division of Psychiatry, Deanery of Clinical Sciences, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh, UK

[9] Salvesen Mindroom Research Centre, University of Edinburgh, Edinburgh, UK

Impact:

- A combination of clinical perinatal factors and neonatal DTI measures of white matter microstructure leads to accurate prediction of language outcome at 2 years corrected gestational age following preterm birth.

- A model that comprises clinical and MRI features that has potential to be scalable across centres. It offers a basis for enhancing the power and generalisability of diagnostic and prognostic studies of neurodevelopmental disorders associated with language impairment.
- Early identification of infants who are at risk of language delay, facilitating targeted early interventions and support services, which could improve the quality of life for children born preterm.

Abstract

Background

Preterm birth can lead to impaired language development. This study aimed to predict language outcomes at two years corrected gestational age (CGA) for children born preterm.

Methods

We analysed data from 89 preterm neonates (median GA 29 weeks) who underwent diffusion MRI (dMRI) at term-equivalent age and language assessment at two years CGA using the Bayley-III. Feature selection and a random forests classifier were used to differentiate typical versus delayed (Bayley-III language composite score<85) language development.

Results

The model achieved balanced accuracy:91%, sensitivity:86%, and specificity:96%. The probability of language delay at two years CGA is increased with: increasing

values of peak width of skeletonised fractional anisotropy (PSFA), radial diffusivity (PSRD), and axial diffusivity (PSAD) derived from dMRI; among twins; and after an incomplete course of, or no exposure to, antenatal corticosteroids. Female sex and breastfeeding during the neonatal period reduced the risk of language delay.

Conclusion

The combination of perinatal clinical information and MRI features leads to accurate prediction of preterm infants who are likely to develop language deficits in early childhood. This model could potentially enable stratification of preterm children at risk of language dysfunction who may benefit from targeted early interventions.

Introduction

An estimated 15 million infants are born preterm (before 37 weeks of gestation) annually worldwide.[1] Although advances in neonatal intensive care have led to a decrease in infant mortality rates over time, survivors of preterm birth are at increased risk of long-term neurocognitive impairment.[2] Preterm birth may lead to language deficits that persist into school age[3] and are associated with a range of negative sequelae across the life span, including poor academic performance, poor social, emotional and behavioural functioning, and unemployment.[4,5] Neurodevelopmental trajectories are amenable to early interventions, which presents a window of opportunity to have a profound, long-lasting effect on later life.[6] Therefore, there is a clear unmet clinical need for early identification of those children who are at high risk of poor language development.

Multiple outcome studies have demonstrated associations between prenatal, neonatal and postnatal factors, and early neurodevelopmental outcomes for preterm infants.[7,8] In addition, preterm birth is closely associated with generalised microstructural changes in cerebral white matter, inferred from diffusion tensor imaging (DTI) (fractional anisotropy [FA], mean, axial, and radial diffusivities [MD, AD, RD]) and alterations in these have been linked to language delay.[9] However, it is rare for research to combine data from different modalities for the development of prediction models for neurodevelopmental outcomes.

Nonetheless, a few studies have built and validated tools for prediction of the composite outcome of neurodevelopmental impairment at 2 years corrected gestational age (CGA) for children born preterm. Tyson et al.[10] investigated the clinical and demographic characteristics of a cohort of infants born before 26 weeks of gestation and found that the risk of adverse neurodevelopmental outcome at 18 to 22 months CGA was predicted using gestational age (GA), sex, exposure to antenatal corticosteroids, multiple birth and birth weight. Ambalavanan et al.[11] reported that neurodevelopmental impairment at 18 to 22 months CGA was predicted by combining sex, respiratory illness severity, and enlarged ventricular size, periventricular leukomalacia or porencephalic cyst on cranial ultrasound. Vesoulis et al.[12] developed a tool for prediction of risk of neurodevelopmental impairment at 18 to 24 months CGA. This tool comprised ventilator days, mode of delivery, exposure to antenatal corticosteroids, retinopathy of prematurity (ROP) requiring surgery, and magnetic resonance imaging (MRI) findings (cerebellar haemorrhage size, cerebellar haemorrhage laterality, intraventricular haemorrhage grade, white matter injury).

However, deficits in different developmental domains require different therapies and targeted support strategies. Thus, tools for stratification of children at high risk of impairment in specific developmental domains would be valuable. Recently, Vassar et al.[13] evaluated the predictive value of structural MRI and DTI variables for classification of very preterm infants at high versus low risk of language delay. They developed a model for prediction of language delay that included DTI variables in three brain regions and achieved 89% sensitivity and 86% specificity. Ball et al.[14] revealed that distinct patterns of brain structure and microstructure following preterm birth are linked to specific clinical and environmental factors, and these patterns correlate with neurodevelopmental outcome at 18 to 24 months CGA. Language outcome was associated with specific neuroanatomic variation, which was linked to: age at scan, need for continuous positive airway pressure, birth weight, GA at birth, parenteral nutrition, surfactant administration, and mechanical ventilation.

In view of this evidence, we hypothesised that a combination of clinical, environmental and imaging factors derived from DTI that capture generalised white matter dysmaturation would potentially enhance the prediction of language outcomes at 2 years CGA following preterm birth. Blesa et al.[15] demonstrated that histogram-based variables derived from DTI (peak width of skeletonised [PS] -FA, -MD, -RD, and -AD), which represent generalised water content and myelination, can be used as biomarkers of microstructural white matter alterations associated with preterm birth. The advantage of the histogram-based framework is that it is fully automated, captures generalised white matter dysmaturation which characterises the encephalopathy of prematurity, is computationally inexpensive compared with tract-specific approaches, and has high inter-scanner reproducibility.[16]

A prediction tool that combines clinical data and imaging biomarkers for early language development is lacking, and yet timely identification of future language deficits has clinical and research implications, because it could stratify infants at most need for early interventions. Here, we aimed to develop a machine learning model that accurately predicts typical versus delayed language outcomes at 2 years CGA using a parsimonious feature set derived from clinical, demographic, and histogram-based variables computed from neonatal brain DTI.

Methods

Participants

Participants were selected from a longitudinal cohort of preterm neonates born at ≤33 weeks of gestation at the Royal Infirmary of Edinburgh between February 2012 and August 2015.[17] Selection from the larger cohort was based on availability of diffusion MRI (dMRI) scans at term-equivalent age and 2-year language outcome. Ethical approval was obtained from the UK National Research Ethics Service (NRES), South East Scotland Research Ethics Committee (NRES numbers 11/55/0061 and 13/SS/0143). Written informed consent from parents/carers was obtained for all neonates. Exclusion criteria for the study were congenital anomalies, chromosomal abnormalities, congenital infections or major overt parenchymal lesions (cystic periventricular leukomalacia, haemorrhagic parenchymal infarction), and post-haemorrhagic ventricular dilatation. Infants with a contraindication to MRI at 3 Tesla were also excluded.

Clinical and Demographic Features

The selection of clinical and demographic features included in models was guided by extant literature linking biological and environmental exposures with neurocognitive development in preterm infants. Specifically, we studied the contribution towards prediction of language outcome at two years CGA of the following features: sex,[10,11,18,19] GA (based on first trimester ultrasound),[10,18] birth weight,[10,20] maternal age,[21] primiparity,[19] twin status,[10,20] maternal Body Mass Index (BMI),[22] medical history of maternal depression,[23] administration of a complete course of antenatal corticosteroids for fetal lung maturation (defined as two doses 24 hours apart), any antenatal corticosteroid exposure,[10,12,19,20] administration of antenatal magnesium sulphate ($MgSO_4$) for neuroprotection,[24] mode of delivery (spontaneous vaginal delivery [SVD] or caesarean section),[19] total days requiring intubation whilst in the Neonatal Intensive Care Unit (NICU),[11,12,18] Bronchopulmonary Dysplasia (BPD, defined as oxygen requirement at ≥36 weeks CGA),[19,20,25,26] late onset sepsis (LOS, defined as blood stream infection occurring ≥72 hours postnatally with (a) bacterial pathogen isolated from blood culture, or (b) blood culture growing coagulase negative staphylococcus, along with one or more signs of generalised infection, and treatment with intravenous antibiotics for 5 or more days),[20] Necrotizing Enterocolitis (NEC, defined as stages two or three according to the modified Bell's staging for NEC[27]),[25,28] ROP treated with laser therapy,[12,29] and type of infant feeding at discharge from the neonatal unit (dichotomised as exclusive maternal breast milk versus exclusive formula or mixed feeding).[30] All infants had placental histopathology performed and histological chorioamnionitis was defined using an established system.[31] Maternal level of education (dichotomised as secondary school or below versus college, university or postgraduate studies),[18–20] and socioeconomic status of the family,

operationalised as Scottish Index of Multiple Deprivation 2016 (SIMD16) quintile, where 1 indicates the most deprived and 5 indicates the least deprived (https://www2.gov.scot/Topics/Statistics/SIMD), were also included.

Image Acquisition

Infants underwent a brain MRI scan at term-equivalent age (38-42 weeks' GA) without sedation, during natural sleep after having been fed and swaddled. Vital signs were monitored throughout the scan, and hearing protection was provided for all neonates (MiniMuffs, Natus). All scans were supervised by a physician and a paediatric nurse trained in neonatal resuscitation.

A Siemens MAGNETOM Verio 3-Tesla MRI clinical scanner (Siemens Healthcare Gmbh, Erlangen, Germany) and 12-channel phased-array head coil were used to acquire dMRI data consisting of 11 T2-and 64 diffusion-weighted (b=750 s/mm^2) single-shot, spin-echo, echo planar imaging volumes collected in the axial plane with 2 mm isotropic voxels (TR=7300 ms, TE= 06 ms, FOV=256 mm, acquired matrix =128×128, 50 contiguous interleaved slices with 2 mm thickness, acquisition time=9 min 29 s).

Image Analysis

For each participant the dMRI was denoised using a Marchenko-Pastur-PCA-based algorithm;[32,33] eddy current and head movement were corrected using outlier replacement[34–36] and bias field inhomogeneity correction was performed by calculating the bias field of the mean b0 volume and applying the correction to all the volumes.[37]

For each participant, PSFA, PSMD, PSRD, PSAD were calculated using age-optimised methods described by Blesa et al.[15] In summary, image data were registered to the Edinburgh Neonatal Atlas$_{50}$ (ENA$_{50}$)[15] using a tensor registration,[38] and their DTI maps were calculated. Subsequently, the individual FA maps were projected into the template skeleton and multiplied by the atlas custom mask. Finally, the peak width of the histogram values within the skeletonised maps was calculated as the difference between the 95th and 5th percentiles.[16] Figure 6.1 illustrates a summary of the process described. The code necessary to calculate histogram-based metrics can be found at https://git.ecdf.ed.ac.uk/jbrl/psmd. Figure 6.2 shows scatterplots of the values of the peak width of skeletonised DTI metrics for all participants.

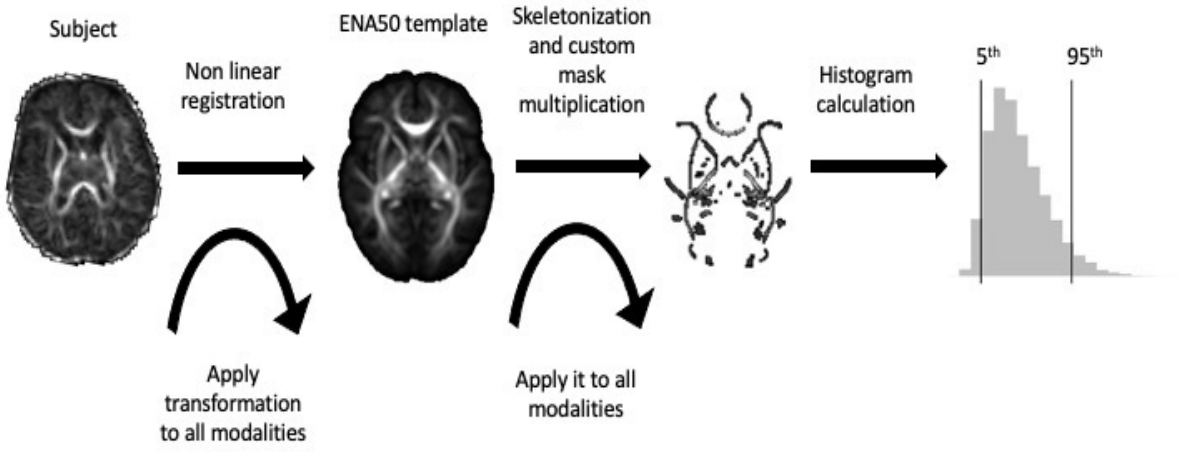


Figure 6.1. Scheme of the steps necessary for the calculation of the peak width of skeletonised DTI metrics. First, participants are registered to a template, then skeletonised and multiplied by a mask to calculate the histogram.

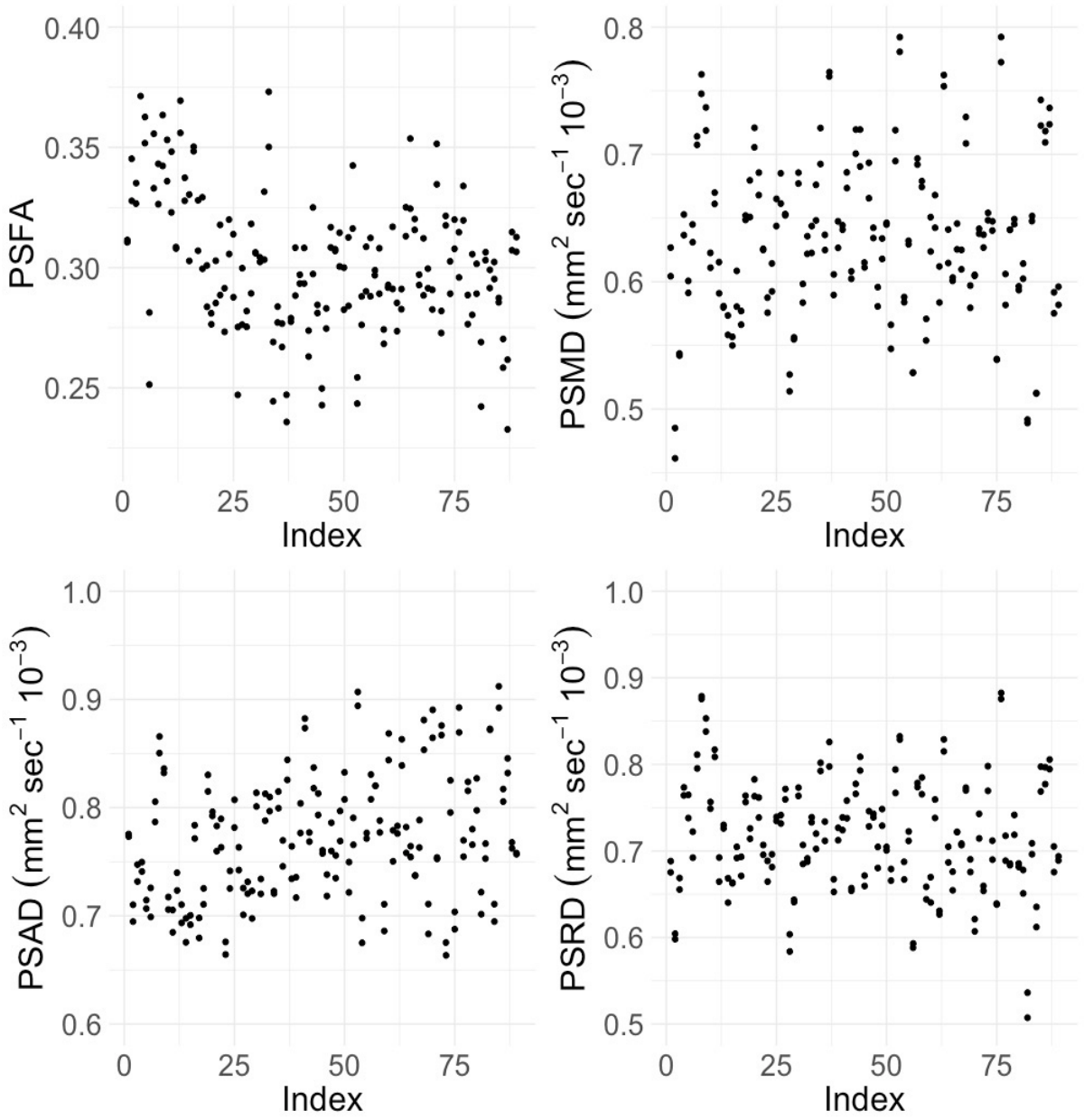


Figure 6.2. Scatterplots of the PSFA, PSMD, PSAD, and PSRD values for all participants. PSFA peak width of skeletonised fractional anisotropy, PSMD peak width of skeletonised mean diffusivity, PSAD peak width of skeletonised axial diffusivity, PSRD peak width of skeletonised radial diffusivity.

Language Outcome

All children took part in a developmental assessment with a trained clinician at 2 years CGA (median age 24.13, range 23.1-28.27 months) using the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III).[39] We used the Bayley-III language composite score (mean 100, SD 15) as the response variable. The clinical cut-off of 85 (i.e. 1 SD below the mean) was used in order to assign children into two distinct groups, thus creating a binary outcome; children whose score was below 85 were considered to have moderate to severe language impairment, while scores equal to or greater than 85 were considered as normal-range or higher.[40]

Data Analysis

We compared three feature selection algorithms: (a) Boruta,[41] (b) ReliefF expRank,[42,43] and (c) Random forests (RF) variable importance.[44] The Boruta algorithm is a wrapper feature selection technique built around the random forests learner, which uses Z score as the importance measure. In other words, it measures the importance of each feature by dividing the average loss of accuracy among all trees by the standard deviation of the accuracy loss. The basic idea of the ReliefF algorithm is to assign a 'weight' value to all features of a dataset based on how well their values distinguish between the instances that are near to each other and thus, how useful they are in predicting the response variable. The important features will have a large weight, while the redundant ones will have a low weight. In random forests variable importance, variable importance is computed using the mean decrease in Gini index. We can measure the total amount that the Gini index is decreased by splits over

a given feature, averaged over all trees. A large value indicates an important feature. In all cases, we obtain a feature ranking indicating in descending order their contribution towards prediction of the response variable. The final feature subset for each feature selection algorithm was selected using Leave-One-Out Cross-Validation (LOOCV), using only the training dataset in each cross-validation iteration and following the process described by Tsanas et al.[45] Subsequently, the selected feature subset was presented into a RF classifier[46] in order to predict the binarised language composite score. Partial Dependence Plots (PDP)[47] were constructed in order to assess how the selected features influence the prediction of the RF classifier. To quantify the strength of the association between the selected features, we used correlation analysis (the Spearman's rank correlation coefficient was used to quantify the strength of the association between two continuous features, the phi coefficient was used to quantify the association between two binary features, and the point-biserial correlation coefficient was used to quantify the strength of the association between a continuous and a binary feature).

The dataset is imbalanced since only 16% of the study group had a language composite score below 85. To overcome the class imbalance problem in the dataset, we explored different data balancing techniques; under-sampling of the majority class, over-sampling of the minority class, and the synthetic minority over-sampling technique (SMOTE),[48] which has been previously used in similar unbalanced applications in the healthcare domain,[49–53] We found that SMOTE yields the best results, which are presented in the paper. SMOTE is a training data enrichment method, where the minority class is over-sampled by creating new synthetic samples, to create a balanced dataset. For each minority class sample, the k minority class nearest neighbours were identified (using the suggestion of Chawla et al. with k=5)

and synthetic samples were introduced along the line segments joining any or all of the k minority class nearest neighbours. Model validation was implemented using LOOCV. LOOCV involves holding out a single observation to be used as the test set, while the learner is trained using the remaining *n*-1 observations (*n* is the total number of observations). The process is repeated *n* times and each time a different observation from the original dataset is used as the test set. The result is *n* estimates of the test error. The final test error rate is the average of these *n* test error estimates. The accuracy of the model was assessed by constructing a confusion matrix which is a contingency table of the observed and predicted classes. Missing data for both numeric and categorical features were imputed using multiple imputation by chained equations (five imputed datasets were created in each LOOCV iteration),[54,55] based only on the information in the training set independently within each LOOCV iteration. Data analysis was conducted in R. The R packages used were: tidyverse, dplyr, caret, randomForest, CORElearn, Boruta, mice, ggplot2, DMwR, Hmisc, RGraphics, grid, gridExtra, gridGraphics.

Results

Two-year language data and dMRI of the brain at term equivalent age were available from 89 children; demographic and clinical characteristics of the study population are presented in Table 6.1. At median age 24.13 months (range 23.24-28.27 months), 14 children had a language composite score below 85. The percentage of missing values in the dataset was 0.2% (one participant had missing histological chorioamnionitis data, two participants had missing SIMD16 and three participants had missing maternal BMI).

Figure 6.3 illustrates the out-of-sample performance of the RF classifier (trained on approximately 150 samples in each LOOCV iteration) as a function of the number of features selected by the different feature selection algorithms. These data show that feeding a subset of eight features selected by the Boruta feature selection algorithm (a wrapper feature selection technique built around the RF learner) to the RF classifier gives the highest balanced accuracy. The selected feature subset comprises PSFA, twin status (yes or no), antenatal steroid exposure (complete or incomplete course), any antenatal steroid exposure (yes or no), sex (male or female), PSRD, PSAD, and feeding at discharge from the NICU (exclusive maternal breast milk versus exclusive formula or mixed feeding). Figure 6.4 shows the importance attributed to each feature by each of the feature selection algorithms. PSFA, twin status, the course of antenatal steroid exposure, any antenatal steroid exposure, sex, PSRD, PSAD, and feeding are the jointly most predictive features towards the prediction of the binarised language outcome. PDP were used to visualise relationships between the selected features and the response based on our model (see Figure 6.5). The PDP provide insight into the effect of changing one or two features in terms of the model's prediction (binary response variable, indicating whether language composite score <85). Regarding the histogram-based variables derived from DTI, the PDP show that the predicted language impairment probability rises with increasing PSFA, PSRD, and PSAD values. PSRD and PSAD are presented in the same plot because they are highly correlated as illustrated in the correlogram and correlation matrix in Figure 6.6. Language composite score <85 at 2 years CGA is more likely following a twin pregnancy, an incomplete course of antenatal corticosteroids, or no exposure to antenatal steroids. Female sex and feeding with exclusive breast milk reduce the risk of future language delay.

Table 6.2 shows the confusion matrix of the out-of-sample classification performance of the RF classifier when mapping the selected feature subset (i.e., PSFA, twin status, antenatal corticosteroid exposure, sex, PSRD, PSAD, and feeding at discharge) to the binarised language composite score. Our model achieved balanced accuracy: 91%, sensitivity: 86%, and specificity: 96%.

Finally, we repeated the analysis to investigate separately the performance of the model when presented only with either clinical or MRI features, which led to reduced model performance. As shown in Table 6.3, the model that comprises clinical and MRI features outperformed the models using only clinical or MRI features. The combination of clinical and DTI features enhances the prediction of language outcomes at 2 years CGA following preterm birth.

Table 6.1. Demographic and clinical characteristics of the study group.

Characteristics	Neonates with language composite score ≥ 85 (N=75)	Neonates with language composite score < 85 (N=14)
Antenatal		
Any antenatal corticosteroids	73 (97)	11 (79)
Complete course of antenatal corticosteroids	56 (75)	5 (36)
Antenatal $MgSO_4$ for fetal neuroprotection	39 (52)	8 (57)
Perinatal		
Sex		
Male	35 (47)	12 (86)
Female	40 (53)	2 (14)
GA (weeks)	28.84±3.28 (23.28-33)	28.92±2.18 (23.28-30.28)
Birth weight (grams)	1137±376.5 (568-1500)	1040±410 (550-1635)
Birth weight z score	-0.16±1.15 (-2.63-1.17)	0.12±1.30 (-1.77-1.0)
Apgar score at 5 minutes	7.5±2 (2-9)	8±2 (5-9)

Mode of delivery		
SVD	32 (43)	3 (21)
Caesarean section	43 (57)	11 (79)
Primiparity	52 (69)	8 (57)
Twin status	21 (28)	10 (71)
Postnatal		
BPD	25 (33)	6 (43)
LOS	20 (27)	5 (36)
NEC	5 (7)	0 (0)
ROP	5 (7)	1 (7)
Histologic chorioamnionitis	22 (31)	3 (21)
Days of intubation	1±5.5 (0-39)	1±1 (0-43)
Feeding at discharge		
Exclusive maternal breast milk	36 (48)	2 (14)
Exclusive formula or mixed feeding	39 (52)	12 (86)
Demographics		
Maternal race		
Asian	5 (6)	0 (0)
White	66 (88)	13 (93)
White/Asian	1 (1)	0 (0)
White/Black	2 (2)	1 (7)
Other mixed	1 (1)	0 (0)
Maternal age (years)	32±8 (17-43)	33±8 (23-40)
Maternal BMI	24.7 ±4.5 (17.4-43)	24.1±6.9 (18-30.9)
Medical history of maternal depression	10 (13)	1 (7)
Maternal education		
Secondary school or below	33 (44)	6 (43)
College/University/postgraduate studies	42 (56)	8 (57)
SIMD16 quintile		
1	8 (11)	3 (21)
2	23 (32)	3 (21)
3	11 (15)	3 (21)
4	12 (16)	3 (21)
5	19 (26)	2 (14)
Histogram-based variables derived from DTI		
PSFA	0.3±0.03 (0.25-0.37)	0.32±0.02 (0.24-0.36)
PSMD	0.63±0.07 (0.49-0.79)	0.61±0.07 (0.49-0.79)
PSRD	0.71±0.09 (0.54-0.88)	0.72±0.07 (0.6-0.83)
PSAD	0.77±0.08 (0.68-0.89)	0.76±0.09 (0.69-0.91)
Bayley-III		
Language composite score	100±24 (86-132)	77±11 (56-83)

Variables are presented in the form median ± IQR (range) or number (%).

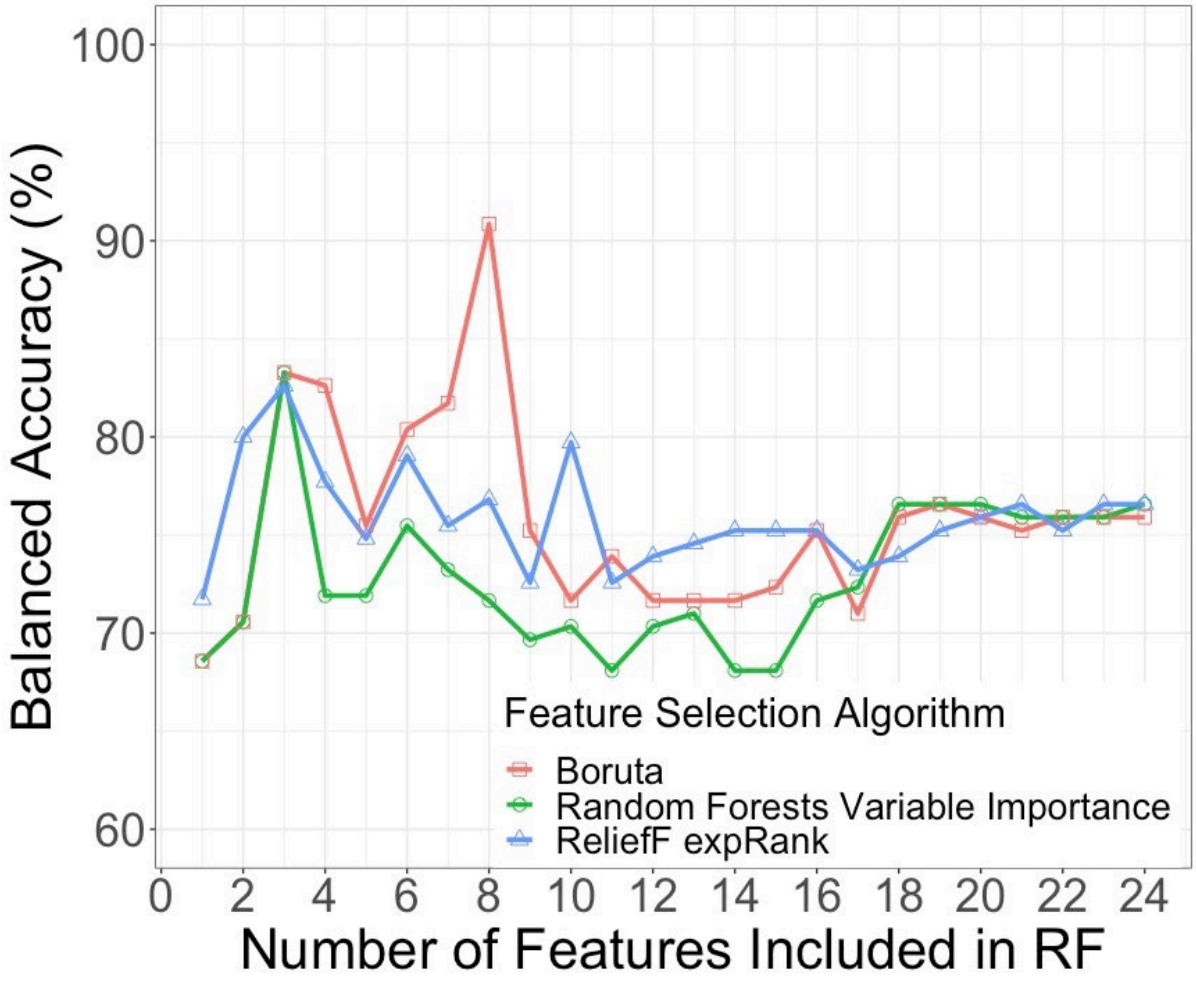


Figure 6.3. Comparison of out-of-sample LOOCV balanced accuracy results of the random forests classifier using the features selected by each of the three feature selection algorithms.

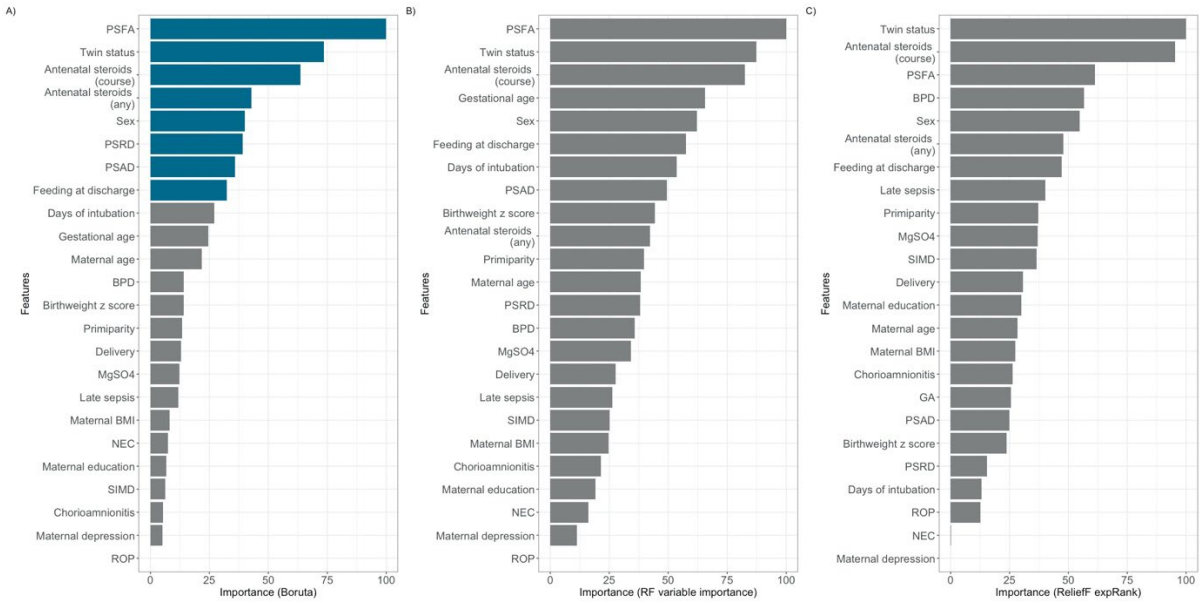


Figure 6.4. Feature importance plots. A) Importance attributed to each feature by the Boruta algorithm. The first eight features coloured in blue (PSFA, twin status, course of antenatal steroids, any antenatal steroids, sex, PSRD, PSAD, feeding at discharge) are the jointly most predictive features towards the prediction of language outcome. **B)** Importance attributed to features by RF variable importance. **C)** Importance attributed to features by ReliefF expRank. Computation of feature importance depends on the feature selection algorithm used and is expressed relative to the maximum.

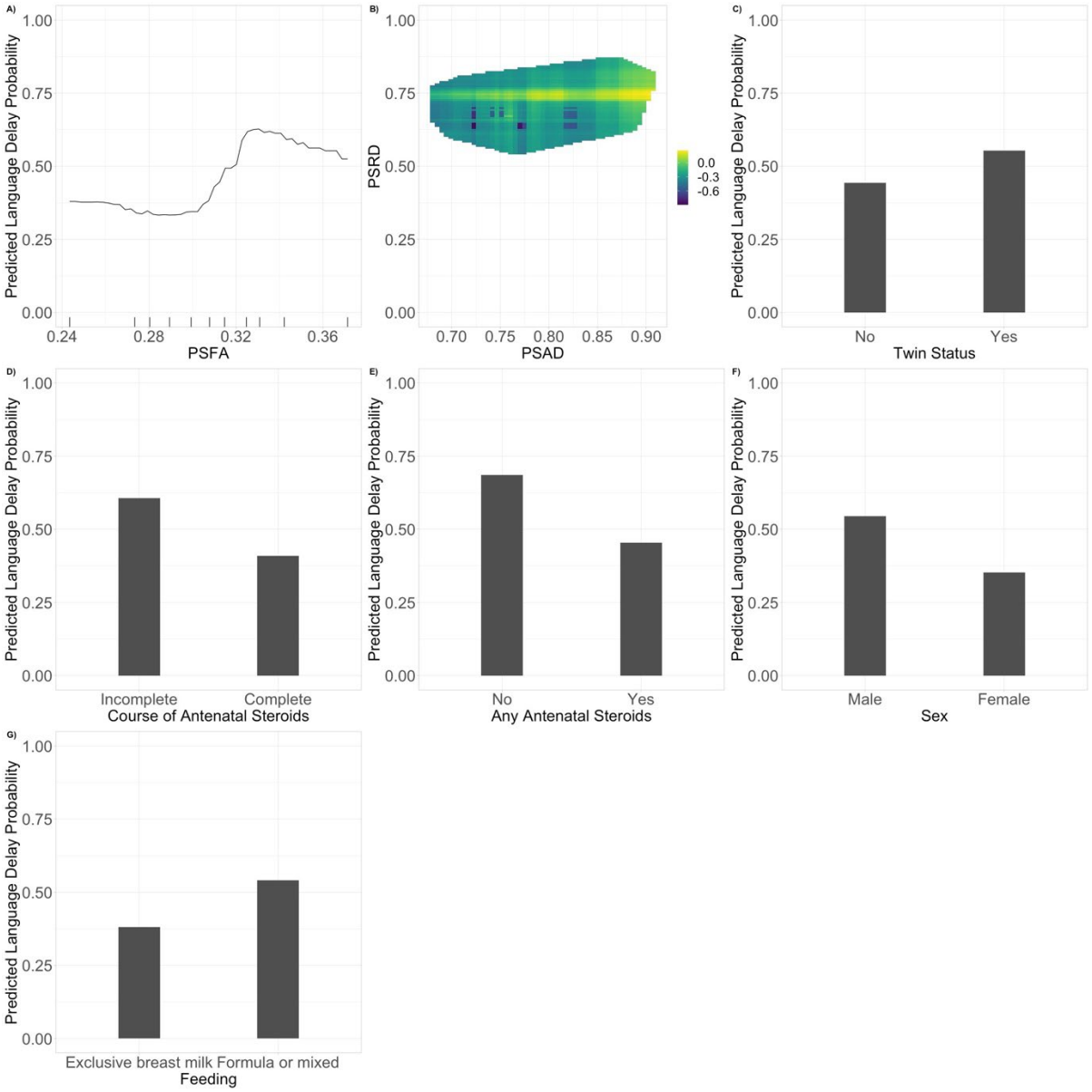
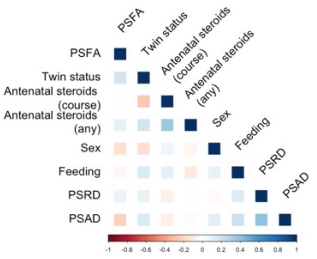


Figure 6.5. Partial dependence plots for the eight features selected by Boruta and used in the random forests classifier. A) The predicted language impairment probability rises with increasing PSFA values. **B)** 3D plot of PSRD and PSAD. The predicted language impairment probability rises with increasing PSRD, and PSAD values. **C)** A twin pregnancy increases the predicted probability of language impairment. **D)** An incomplete course of antenatal corticosteroids increases the predicted probability of language impairment. **E)** No exposure to any antenatal steroids increases the predicted probability of language impairment. **F)** Female sex reduces the predicted probability of language impairment. **G)** Feeding with exclusive breast milk reduce the predicted probability of language impairment. Language composite score <85 at 2 years CGA is more likely following a twin pregnancy, an incomplete course of antenatal corticosteroids, or no exposure to any antenatal steroids. Female sex and feeding with exclusive breast milk reduce the risk of future language delay.



	PSFA	Twin status	Antenatal steroids (course)	Antenatal steroids (any)	Sex	PSRD	PSAD
PSFA							
Twin status	0.22*						
Antenatal steroids (course)	0.03	-0.27*					
Antenatal steroids (any)	0.13	0.18	0.36***				
Sex	-0.19	-0.17	0.06	-0.06			
PSRD	0.09	0.08	-0.07	-0.05	-0.03		
PSAD	-0.24*	0.14	-0.06	-0.05	0.13	0.41****	
Feeding	-0.03	0.15	0.10	-0.11	0.09	0.16	0.21*

Figure 6.6. Correlogram and correlation matrix of the eight most important features selected by the Boruta algorithm. p<.05 '*', p<.01 '**', p<.001 '***', p<.0001 '****'.

Table 6.2. Confusion matrix summarizing the out-of-sample findings using LOOCV.

Prediction	**Reference**	
	Language composite score <85	**Language composite score ≥85**
Language composite score <85	12 (14%)	3 (3%)
Language composite score ≥85	2 (2%)	72 (81%)

Table 6.3. Model performance using (a) only clinical features, (b) only MRI features, and (c) the combination of clinical and MRI features.

Models	**Balanced accuracy**	**Sensitivity**	**Specificity**
Clinical features	83%	79%	87%
MRI features	81%	86%	76%
Clinical and DTI features	91%	86%	96%

Discussion

We developed a parsimonious machine learning model which accurately identifies preterm infants who are likely to develop language impairment in early childhood. We explored the predictive value of 24 clinical, demographic and brain imaging features, and found that a robust subset of eight clinical characteristics and imaging biomarkers best predicts a language composite score below 85 on the Bayley-III: PSFA, PSRD, PSAD, twin status, administration of an incomplete course of antenatal corticosteroids, no exposure to antenatal corticosteroids, male sex, and feeding with exclusive formula milk or mixed formula and breast milk. Overall, we demonstrated out-of-sample balanced accuracy: 91%, sensitivity: 86%, and specificity: 96%.

Feature selection was conducted by comparing three feature selection algorithms: (a) Boruta, (b) ReliefF expRank, and (c) RF variable importance. Feature selection methods can be broadly considered into three main categories: filter, wrapper, embedded methods. Filter feature selection methods work independently of a statistical learner relying on the general statistical properties of the data, and thus select a feature subset which is not tuned or optimised towards a specific learning algorithm. Wrapper methods take a particular machine learning method into account in order to choose the best subset of the original features. They evaluate multiple models by training and testing in the feature space, thus optimizing the performance of the particular machine learning model that was used. Embedded methods choose the subset of features while the learning model is being constructed. This means that the resulting feature subset is specific to a particular learning algorithm. We chose to use a feature selection algorithm from each main category for our exploration; ReliefF is a filter technique, Boruta is wrapper feature selection technique built around the

random forests learner, and RF variable importance is an embedded method. The use of ReliefF and the RF importance have been extensively used and validated in many different applications and we have previously conducted a thorough empirical study[56] where they performed very competitively against many established feature selection approaches. In general, we would expect a wrapper or embedded method to perform better for a particular choice of a classifier, although it might not necessarily generalise very well with the choice of different classifiers."

Our findings suggest that PSFA, PSRD, and PSAD, which detect generalised white matter microstructural alterations in preterm infants compared to infants born at term,[15] are predictive of impaired language development at two years CGA. We explored the predictive value of whole brain measures of peak width of skeletonised DTI metrics, instead of tract specific segmentations, because preterm brain dysmaturation is a substantially generalised process,[57] and language development draws on broad cognitive capacities. We have found that the probability of language delay is higher with increased PSFA, PSRD, and PSAD. These features are consistent with delayed myelination, less coherent white matter organisation, and altered axonal integrity in the preterm brain.[15,58] Previous research has also shown that abnormalities in brain structure following preterm birth are correlated with long-term neurodevelopmental outcome.[59]

The data show that twin status is associated with increased risk of impaired language development. This finding is consistent with studies in the extant literature which have found that multiple pregnancy is associated with neurodevelopmental impairment[10,20,60] and language delay[61] at 2 years CGA. Language delay in twins can

be attributed to postnatal environmental factors;[62,63] twins receive a less focused and less elaborated communicative interchange with their parents than do singletons. Thorpe et al.[62] compared families with twins to families with pairs of closely spaced singletons. This study found that language delay in twins compared to singletons may be explained by patterns of parent-child interaction and communication.

Moreover, previous work has shown that exclusive breast milk feeding in the weeks following preterm birth can enhance brain development,[30] and in the general population breast milk intake in infancy is associated with improved performance on intelligence tests.[64] In line with this, we found that exclusive breastfeeding is associated with improved language outcomes compared to formula feeding or mixed breast and formula feeding. It is surprising that GA at birth was not included in the final feature set. However, its influence on long-term outcome may be captured by PSRD and PSAD which are strongly correlated with GA at birth.[15]

This study is the first to investigate the use of peak width of skeletonised DTI metrics as predictors for language development in the preterm population. The advantage of using these image biomarkers is that their calculation is fully automated, computationally inexpensive, and has high inter-scanner reproducibility,[16] meaning that they can be easily obtained for preterm neonates who undergo a dMRI scan at term-equivalent age and can be used for multi-centre studies. Thus, our model comprises features which can be easily obtained for future clinical application.

Hitherto, few studies have focused on developing and validating prediction models for early neurodevelopmental outcomes for children born preterm. Most tools predict the

composite outcome of neurodevelopmental impairment.[10–12] However, deficits in different developmental domains require different interventions. Therefore, tools for timely identification of children at risk of impairment in specific developmental domains are valuable. The developed model predicts language deficits at 2 years CGA. Recently, a model was developed for classification of very preterm infants at high versus low risk for language delay, which achieved 89% sensitivity and 86% specificity.[13] That model included DTI variables in three brain regions: MD of right sagittal stratum and right inferior occipital gyrus, and AD of right lingual gyrus. However, whole brain calculation of DTI variables is computationally expensive, hence we investigated the predictive value of histogram-based variables derived from DTI. We have shown that combining DTI metrics with perinatal factors, along with the use of advanced machine learning techniques can further improve identification of children at risk of language impairment.

The main strength of our study is that we had a longitudinal cohort of preterm infants that is deeply phenotyped with brain imaging and biological information that enabled us to investigate a large number of clinical, demographic, social, and DTI variables. We acknowledge some limitations in our study. The sample size is relatively small, and this is a single centre study so despite our best efforts with standard model validation techniques to assess model generalisation we would need to further validate findings in a different cohort. Nonetheless, the study population was fairly representative of NICU populations in terms of comorbidities that have been associated with long-term neurodevelopmental outcomes. In addition, cortical grey matter was not assessed in this study. We focused on alterations in white matter microstructure, since it is the most consistently abnormal finding in preterm infants, by

measuring a functionally tractable property using a tool that is readily applied to clinical image data. Future studies could aim to validate our model in additional external cohorts, and also apply machine learning techniques for prediction of motor, cognitive and social-emotional outcomes for children born preterm.

Conclusion

A combination of clinical perinatal factors and neonatal DTI measures of white matter microstructure best predict language impairment at 2 years after preterm birth. This model has the potential to enable clinicians identify infants who are at risk of language delay, thus facilitating targeted early intervention and support services. The model comprises clinical and MRI features that have potential to be scalable across centres, so it offers a basis for enhancing the power and generalisability of diagnostic and prognostic studies of neurodevelopmental disorders associated with language impairment.

References

1. Chawanpaiboon S, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob. Health.* 2019;7:e37–e46.

2. Pierrat V, et al. Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in France in 2011: EPIPAGE-2 cohort study. *BMJ.* 2017;358:j3448.

3. van Noort-van der Spek IL, Franken M-CJP, Weisglas-Kuperus N. Language functions in preterm-born children: a systematic review and meta-analysis. *Pediatrics.* 2012;129:745–754.

4. Law J, Rush R, Schoon I, Parsons S. Modeling developmental language difficulties from school entry into adulthood: literacy, mental health, and employment outcomes. *J. Speech Lang. Hear. Res.* 2009;52:1401–1416.

5. Conti-Ramsden G, Mok PLH, Pickles A, Durkin K. Adolescents with a history of specific language impairment (SLI): strengths and difficulties in social, emotional and behavioral functioning. *Res. Dev. Disabil.* 2013;34:4161–4169.

6. Spittle, A., Orton, J., Anderson, P. J., Boyd, R. & Doyle, L. W. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database Syst. Rev*. CD005495 (2015).

7. Linsell L, Malouf R, Morris J, Kurinczuk JJ, Marlow N. Prognostic factors for poor cognitive development in children born very preterm or with very low birth weight: a systematic review. *JAMA Pediatr.* 2015;169:1162–1172.

8. Linsell L, Malouf R, Morris J, Kurinczuk JJ, Marlow N. Prognostic factors for cerebral palsy and motor impairment in children born very preterm or very low birthweight: a systematic review. *Dev. Med. Child Neurol.* 2016;58:554–569.

9. Feldman HM, Lee ES, Yeatman JD, Yeom KW. Language and reading skills in school-aged children and adolescents born preterm are associated with white matter properties on diffusion tensor imaging. *Neuropsychologia.* 2012;50:3348–3362.

10. Tyson JE, et al. Intensive care for extreme prematurity–moving beyond gestational age. *N. Engl. J. Med.* 2008;358:1672–1681.

11. Ambalavanan N, et al. Outcome trajectories in extremely preterm infants. *Pediatrics.* 2012;130:e115–e125.

12. Vesoulis, Z. A., El Ters, N. M., Herco, M., Whitehead, H. V & Mathur, A. M. A web-based calculator for the prediction of severe neurodevelopmental impairment in preterm infants using clinical and imaging characteristics. *Children***5**, 151 (2018).
13. Vassar, R. et al. Neonatal brain microstructure and machine-learning-based prediction of early language development in children born very preterm. *Pediatr. Neurol*. **108**, 86–92 (2020).
14. Ball G, et al. Multimodal image analysis of clinical influences on preterm brain development. *Ann. Neurol.* 2017;82:233–246.
15. Blesa M, et al. Peak width of skeletonized water diffusion MRI in the neonatal brain. *Front. Neurol.* 2020;11:235.
16. Baykara E, et al. A novel imaging marker for small vessel disease based on skeletonization of white matter tracts and diffusion histograms. *Ann. Neurol.* 2016;80:581–592.
17. Boardman JP, et al. Impact of preterm birth on brain development and long-term outcome: protocol for a cohort study in Scotland. *BMJ Open.* 2020;10:e035854.
18. Charkaluk ML, et al. Neurodevelopment of children born very preterm and free of severe disabilities: the Nord-Pas de Calais Epipage cohort study. *Acta Paediatr.* 2010;99:684–689.
19. Wood NS, et al. The EPICure study: associations and antecedents of neurological and developmental disability at 30 months of age following extremely preterm birth. *Arch. Dis. Child. Fetal Neonatal Ed.* 2005;90:F134–F140.
20. Vohr BR, Wright LL, Poole WK, McDonald SA. Neurodevelopmental outcomes of extremely low birth weight infants <32 weeks' gestation between 1993 and 1998. *Pediatrics.* 2005;116:635–643.
21. Tseng K-T, et al. The impact of advanced maternal age on the outcomes of very low birth weight preterm infants. *Medicine.* 2019;98:e14336.
22. Reynolds LC, Inder TE, Neil JJ, Pineda RG, Rogers CE. Maternal obesity and increased risk for autism and developmental delay among very preterm infants. *J. Perinatol.* 2014;34:688–692.
23. Bozkurt O, et al. Does maternal psychological distress affect neurodevelopmental outcomes of preterm infants at a gestational age of ≤32weeks. *Early Hum. Dev.* 2017;104:27–31.
24. Marret S, et al. [Effect of magnesium sulphate on mortality and neurologic morbidity of the very-preterm newborn (of less than 33 weeks) with two-year

neurological outcome: results of the prospective PREMAG trial] *Gynecol. Obstet. Fertil.* 2008;36:278–288.

25. Synnes A, et al. Determinants of developmental outcomes in a very preterm Canadian cohort. *Arch. Dis. Child. Fetal Neonatal Ed.* 2017;102:F235–F234.

26. Twilhaar ES, et al. Cognitive outcomes of children born extremely or very preterm since the 1990s and associated risk factors: a meta-analysis and meta-regression. *JAMA Pediatr.* 2018;172:361–367.

27. Bell MJ, et al. Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. *Ann. Surg.* 1978;187:1–7.

28. van Vliet EOG, de Kieviet JF, Oosterlaan J, van Elburg RM. Perinatal infections and neurodevelopmental outcome in very preterm and very low-birth-weight infants: a meta-analysis. *JAMA Pediatr.* 2013;167:662–668.

29. Schmidt B, Davis PG, Asztalos EV, Solimano A, Roberts RS. Association between severe retinopathy of prematurity and nonvisual disabilities at age 5 years. *JAMA.* 2014;311:523.

30. Blesa M, et al. Early breast milk exposure modifies brain connectivity in preterm infants. *Neuroimage.* 2019;184:431–439.

31. Anblagan D, et al. Association between preterm brain injury and exposure to chorioamnionitis during fetal life. *Sci. Rep.* 2016;6:37932.

32. Veraart J, Fieremans E, Novikov DS. Diffusion MRI noise mapping using random matrix theory. *Magn. Reson. Med.* 2016;76:1582–1593.

33. Tournier J-D, et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage.* 2019;202:116137.

34. Smith SM, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage.* 2004;23:S208–S219.

35. Andersson JLR, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage.* 2016;125:1063–1078.

36. Andersson JLR, Graham MS, Zsoldos E, Sotiropoulos SN. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. *Neuroimage.* 2016;141:556–572.

37. Tustison NJ, et al. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging.* 2010;29:1310–1320.

38. Zhang H, Yushkevich PA, Alexander DC, Gee JC. Deformable registration of diffusion tensor MR images with explicit orientation optimization. *Med. Image Anal.* 2006;10:764–785.
39. Albers CA, Grieve AJ. Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development–Third Edition. San Antonio, TX: Harcourt Assessment. *J. Psychoeduc. Assess.* 2007;25:180–190.
40. Johnson S, Moore T, Marlow N. Using the Bayley-III to assess neurodevelopmental delay: which cut-off should be used? *Pediatr. Res.* 2014;75:670–674.
41. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J. Stat. Softw.* 2010;36:1–13.
42. Kira, K. & Rendell, L. A. *The Feature Selection Problem: Traditional Methods and a New Algorithm* (AAAI Press, 1992).
43. Kononenko, I. In *Machine Learning: ECML-94. ECML 1994. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, Vol. 784 (eds Bergadano, F. & De Raedt, L.) 171–182 (Springer, 1994).
44. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*(Springer, 2009).
45. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Trans. Biomed. Eng.* 2012;59:1264–1271.
46. Breiman L. Random forests. *Mach. Learn.* 2001;45:5–32.
47. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 2001;29:1189–1232.
48. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 2002;16:321–357.
49. Dessie EY, Tsai JJP, Chang J-G, Ng K-L. A novel miRNA-based classification model of risks and stages for clear cell renal cell carcinoma patients. *BMC Bioinformatics.* 2021;22:270.
50. Park KH, Batbaatar E, Piao Y, Theera-Umpon N, Ryu KH. Deep learning feature extraction approach for hematopoietic cancer subtype classification. *Int. J. Environ. Res. Public Health.* 2021;18:1–24.
51. Lee, Y. W., Choi, J. W. & Shin, E. H. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med*. **129**, 104151 (2021).

52. Ivanović, M. D. et al. Predicting defibrillation success in out-of-hospital cardiac arrested patients: Moving beyond feature design. *Artif. Intell. Med*. **110**, 101963 (2020).
53. Nguyen QDN, Liu AB, Lin CW. Development of a neurodegenerative disease gait classification algorithm using multiscale sample entropy and machine learning classifiers. *Entropy.* 2020;22:1340.
54. Raghunathan, T. E., Lepkowski, J., Hoewyk, J., Van & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.***27**, 85–95 (2001).
55. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* 2007;16:219–242.
56. Tsanas, A. *Accurate Telemonitoring of Parkinson's Disease Symptom Severity using Nonlinear Speech Signal Processing and Statistical Machine Learning*. PhD thesis, Oxford Univ. (2012).
57. Telford EJ, et al. A latent measure explains substantial variance in white matter microstructure across the newborn human brain. *Brain Struct. Funct.* 2017;222:4023–4033.
58. Boardman JP, Counsell SJ. Invited Review: Factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. *Neuropathol. Appl. Neurobiol.* 2020;46:413–421.
59. Batalle D, Edwards AD, O'Muircheartaigh J. Annual Research Review: Not just a small adult brain: understanding later neurodevelopment through imaging the neonatal brain. *J. Child Psychol. Psychiatry Allied Discip.* 2018;59:350–371.
60. Wadhawan R, et al. Twin gestation and neurodevelopmental outcome in extremely low birth weight infants. *Pediatrics.* 2009;123:e220–e227.
61. Adams-Chapman, I., Bann, C. M., Vaucher, Y. E., Stoll, B. J. & Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Association between feeding difficulties and language delay in preterm infants using Bayley Scales of Infant Development-Third Edition. *J. Pediatr*. **163**, 680.e3–685.e3 (2013).
62. Thorpe K, Rutter M, Greenwood R. Twins as a natural experiment to study the causes of mild language delay: II: Family interaction risk factors. *J. Child Psychol. Psychiatry.* 2003;44:342–355.

63. Thorpe K. Twin children's language development. *Early Hum. Dev.* 2006;82:387–395.
64. Horta BL, Loret De Mola C, Victora CG. Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta Paediatr.* 2015;104:14–

Author Contributions: All authors have made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data. All authors have drafted the article or revised it critically for important intellectual content and have approved the final version to be considered for publication.

Funding information: This work was supported by Theirworld (www.theirworld.org). The work was undertaken in the MRC Centre for Reproductive Health, which was funded by MRC Centre Grant (MRC G1002033). The study was also supported by Health Data Research UK which receives its funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and the Welcome Trust. The funders had no role in the study and the decision to submit this work to be considered for publication.

Competing interests: The authors declare no competing interests.

Ethics approval and consent to participate: Ethical approval was obtained from the UK National Research Ethics Service (NRES), South East Scotland Research Ethics Committee (NRES numbers 11/55/0061 and 13/SS/0143). Written informed consent from parents/carers was obtained for all neonates.

6.3. Conclusion

In this article, we focused on language development following preterm birth. In line with our initial hypothesis, we showed that a combination of clinical perinatal factors and neonatal DTI measures that capture diffuse white matter microstructure alterations can accurately predict language impairment at two years CGA following preterm birth. Specifically, we found that the risk of language impairment (defined as Bayley-III language composite score lower than 85) rises with increasing values of peak width of skeletonised fractional anisotropy (PSFA), radial diffusivity (PSRD), and axial diffusivity (PSAD) derived from dMRI. In addition, language impairment is more likely following a twin pregnancy, an incomplete course of antenatal corticosteroids, or no exposure to antenatal corticosteroids. On the contrary, female sex and feeding with exclusive breast milk reduce the risk of language delay. Our model achieved balanced accuracy: 91%, sensitivity: 86%, and specificity: 96%, outperforming existing statistical models.

This study has contributed to our broader research goal by developing a model that accurately predicts language outcomes at two years CGA following preterm birth. The proposed tool has the potential to timely identify preterm infants at high risk of language impairment, who may benefit from targeted early interventions and support services. Early interventions can help preterm infants reach their full developmental potential, thus improving the quality of life for this vulnerable population. Moreover, this study, in line with the study presented in Chapter 5, has highlighted the importance of perinatal practices, such as breast milk feeding, for preterm brain development and subsequent neurodevelopmental outcomes.

CHAPTER 7: DISCUSSION

7.1. Summary of findings

This thesis investigated the combination of early life exposures that significantly impact brain development and subsequent neurodevelopmental outcomes following preterm birth. The primary objectives were to identify potentially modifiable risk factors that influence brain growth and to predict which preterm infants are at the greatest risk of developing long-term neurodevelopmental delays, with a focus on language outcomes. The underlying rationale was to guide the improvement of perinatal treatments and interventions and to enable the timely identification of preterm infants who may benefit from early support services.

Our study on the prediction of global and regional brain volumes revealed that a combination of clinical and environmental exposures best predicts brain growth in preterm infants. In line with our initial hypothesis, we found that different brain regions are differentially vulnerable to perinatal exposures. Specifically, a robust subset of six features (birthweight z-score, GA at birth, GA at MRI scan, sex, duration of intubation while in the NICU, and socioeconomic status operationalised as neighbourhood deprivation) best predicts total brain volume, with RMSE = 29.23cm^3, MAE = 23.89cm^3, and MAPE = 6.57%. Birthweight z-score, GA at birth and MRI scan, and sex are the selected feature set towards the prediction of white matter volume in preterm infants (RMSE = 14.92cm^3, MAE = 11.98cm^3, MAPE = 7.16%). Cortical grey matter volume is predicted by the combination of birthweight z-score, GA at birth, and

GA at MRI scan (RMSE = 14.43cm^3, MAE = 11.45cm^3, MAPE = 8.47%). The major predictors of deep nuclear grey matter volume are birthweight z-score, GA at birth, GA at MRI scan, sex, duration of intubation and exposure to breast milk feeding while in the NICU (RMSE = 1.92cm^3, MAE = 1.45cm^3, MAPE = 5.54%). Finally, cerebellar volume at term-equivalent age is predicted by the combination of birthweight z-score, GA at birth and MRI scan, sex, socioeconomic deprivation, and duration of parenteral nutrition (RMSE = 3.08cm^3, MAE = 2.18cm^3, MAPE = 8.75%). Overall, this study demonstrated that potentially modifiable risk factors, including postnatal nutrition, respiratory morbidity, and socioeconomic status of the family, significantly influence brain development and subsequent neurodevelopment in preterm infants.

Furthermore, neurodevelopmental outcomes can potentially be enhanced by targeted early interventions and the provision of developmental support. To this end, we developed a machine learning model which accurately predicts language impairment at 2 years CGA following preterm birth. Our model outperformed existing statistical models, achieving a balanced accuracy of 91%, with a sensitivity of 86%, and a specificity of 96%. The model comprised clinical features that are collected as part of routine care and imaging features which capture diffuse white matter microstructural changes in the brain and are readily calculable from dMRI. We found that the probability of poor language development increases with higher values of PSFA, PSRD, and PSAD, which indicate delayed myelination, less coherent white matter organisation, and altered axonal integrity in the preterm brain. Additional risk factors include twin pregnancy, an incomplete course or no exposure to antenatal corticosteroids, male sex, and feeding with formula milk or mixed formula and breast milk.

In summary, the research we undertook has identified the early life risk factors that predict brain growth and language impairment following preterm birth.

7.2. Limitations of the study

Although our study has provided valuable insights into the early life risk factors that impact brain development, as well as the clinical perinatal factors and imaging biomarkers that jointly predict language outcomes following preterm birth, it is important to acknowledge several limitations. Firstly, the sample size in both studies described in Chapters 5 and 6 was relatively small for machine learning analysis. Additionally, some features, such as NEC and ROP, may have not been selected as important predictors by the feature selection algorithms due to their low frequency in our sample (see Tables 5.1 and 6.1 in Chapters 5 and 6, respectively). Moreover, we did not assess the predictive value of cortical grey matter pathology for language impairment at two years CGA (see Chapter 6). Furthermore, we did not have access to an external cohort to replicate our findings and objectively assess the wider generalisation of the developed models. Instead, we assessed model generalisation using standard model validation techniques, such as cross-validation. Finally, we need to acknowledge the inevitable impact the COVID-19 pandemic has had on our study; the main challenge we faced was the delay in data collection regarding the 2-year appointments for assessment of preterm infants with the Bayley-III scales, which resulted in significantly smaller sample sizes than initially expected.

7.3. Implications for practice

To consider the implications for practice of our research, let us assume that our results are robust and replicable. The findings of our first study on the prediction of brain volumes following preterm birth (see Chapter 5) have shown that birthweight z-score, GA at birth, age at MRI scan, and sex, in combination with potentially modifiable risk factors, including postnatal nutrition, respiratory morbidity, and socioeconomic status of the family affect brain growth following preterm birth. Identifying the most important perinatal risk factors that impact brain development in preterm infants is crucial in order to inform current perinatal practices. Our findings suggest that neuroprotective strategies and preventive interventions should aim to promote breast milk nutrition during NICU hospitalisation, utilise ventilation strategies that minimise respiratory morbidity, and address family socioeconomic deprivation. By targeting these risk factors, health professionals can offer preterm infants a healthier start in life, and thus improved long-term outcomes and a better quality of life.

Our second study focused on the prediction of language outcomes at two years CGA following preterm birth (see Chapter 6). The model we have developed comprises features which can be easily obtained in everyday clinical practice, including dMRI data, type of postnatal nutrition, sex, exposure to antenatal corticosteroids, and twin pregnancy. Hence, this model has the potential to be used in clinical practice, enabling clinicians to timely detect which preterm infants are at high risk of future language deficits. Early identification can allow for targeted early interventions and support services which can enhance long-term outcomes.

7.4. Implications for future research

Based on the findings of this study and the limitations outlined in the preceding sections, we suggest several recommendations for further research. First, given that this was a single-centre study in a tertiary hospital in a developed country, future research could aim to replicate our study in diverse cohorts to assess the generalisability of our findings across different populations and settings. Moreover, given the relatively small sample size of our study, future studies could aim to validate our machine learning models in external cohorts to confirm the robustness of our findings.

Furthermore, future research could build on our work by evaluating the predictive importance of additional features that could be used as early biomarkers of brain development and subsequent neurodevelopmental outcomes following preterm birth. Variables such as chorioamnionitis, funisitis, and villitis, which we did not investigate in our study on prediction of cerebral volumes at term-equivalent age, could potentially improve the performance of our models. Moreover, in this study, we focused on neighbourhood deprivation as a proxy for assessing the socioeconomic position of the family. Future studies could investigate in more detail how the social determinants of health, including parents' educational level, income, living and working conditions, and access to healthcare affect brain growth and subsequent neurodevelopmental outcomes. Moreover, eye-tracking which assesses eye-gaze behaviour in response to visual stimuli, permits inference about underlying cognitive processes(Brady et al., 2014; Finke et al., 2017; Gillespie-Smith et al., 2016; Telford et al., 2016). Future

studies could explore whether eye-tracking could enhance the prediction of neurodevelopmental deficits following preterm birth.

This study focused on predicting language outcomes at two years CGA following preterm birth. However, as already described, deficits in different developmental domains require different interventions. With our research, we have optimised the methodology for the prediction of neurodevelopmental outcomes following preterm birth. Hence, future studies could use the same methodology to develop machine learning models for the prediction of a range of neurodevelopmental outcomes, including cognitive, motor, adaptive functioning, and social-emotional outcomes at two years CGA for children born preterm. Timely identification of deficits in different developmental domains will assist in targeting intervention programmes and prioritizing skill areas for interventions to prevent or reduce later difficulties and provide the best opportunity for improved development and better quality of life. Finally, it would be beneficial to develop models comprising early life risk factors to predict neurodevelopmental outcomes at older ages, such as at five years.

7.5. Conclusion

In conclusion, this thesis has demonstrated that a combination of clinical and environmental exposures significantly impacts brain growth following preterm birth. Addressing the key factors identified in this research, including postnatal nutrition, ventilation strategies in the NICU, and socioeconomic deprivation, can potentially enhance brain development and subsequent neurodevelopmental outcomes of preterm infants. Moreover, our findings indicate that early life risk factors, including

clinical perinatal and imaging features at term-equivalent age, can accurately predict language impairment at two years CGA. This means that clinicians can identify high-risk infants even during hospitalisation in the NICU, allowing for targeted interventions and support services in early childhood which can potentially improve long-term outcomes. Overall, our research provides health professionals with the opportunity to treat and intervene early, supporting preterm infants in achieving their full developmental potential and enjoying a better quality of life.

CONFERENCE PROCEEDINGS

- **Evdoxia Valavani,** Dimitrios Doudesis, Ioannis Kourtesis, Richard F. M. Chin, Donald J. MacIntyre, Sue Fletcher-Watson, James P. Boardman, Athanasios Tsanas. Data-Driven Insights Towards Risk Assessment of Postpartum Depression. BIOSTEC 2020, Valletta, Malta, 2020 (oral presentation)

- **Evdoxia Valavani**, Manuel Blesa, Paola Galdi, Gemma Sullivan, Bethan Dean, Hilary Cruickshank, Magdalena Sitko-Rudnicka, Mark E Bastin, Richard F M Chin, Donal J MacIntyre, Sue Fletcher-Watson, James P Boardman, Athanasios Tsanas. Machine learning for stratification of children at risk of language delay following preterm birth. PAS 2021 VIRTUAL (e-poster)

- **Evdoxia Valavani,** Manuel Blesa, Paola Galdi, Gemma Sullivan, Bethan Dean, Hilary Cruickshank, Magdalena Sitko-Rudnicka, Mark E Bastin, Richard F M Chin, Donal J MacIntyre, Sue Fletcher-Watson, James P Boardman, Athanasios Tsanas. 60th Panhellenic Congress of Paediatrics, Chalkidiki, 2021 (oral presentation)

- **Evdoxia Valavani**, Manuel Blesa, Katie Mckinnon, Kadi Vaher, Gemma Sullivan, Mark Bastin, Michael thrippleton, Richard Chin, Donald MacIntyre,

James Boardman, Athanasios Tsanas. Early Life Determinants of Brain Growth Following Preterm Birth: Prediction Using Machine Learning. OHBM 2023, Montréal, Canada, 2023 (poster)

PUBLISHED PEER-REVIEWED JOURNAL ARTICLES

- **Valavani Evdoxia**, Manuel Blesa, Paola Galdi, Gemma Sullivan, Bethan Dean, Hilary Cruickshank, Magdalena Sitko-Rudnicka, Mark E Bastin, Richard F M Chin, Donal J MacIntyre, Sue Fletcher-Watson, James P Boardman, Athanasios Tsanas. Language function following preterm birth: prediction using machine learning. *Paediatric Research,* 2022 Aug;92(2):480-489

- Katie McKinnon, Paola Galdi, Manuel Blesa-Cabez, Gemma Sullivan, Kadi Vaher, Amy Corrigan, Jill Hall, Lorena Jimenez-Sanchez, Michael Thrippleton, Mark E Bastin, Alan J Quigley, **Evdoxia Valavani**, Athanasios Tsanas, Hilary Richardson, James P Boardman. Association of Preterm Birth and Socioeconomic Status With Neonatal Brain Structure. *JAMA Network Open,* 2023 May 1;6(5):e2316067

AWARDS

- Student Research Award 2021, Society for Paediatric Research
  For the study “Machine learning for stratification of children at risk of language delay following preterm birth” presented at the Paediatric Academic Societies (PAS) Meeting 2021

- Trainee Registration Grant, Paediatric Academic Societies (PAS) 2021

REFERENCES

Albers, C. A., & Grieve, A. J. (2007). Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development– Third Edition. San Antonio, TX: Harcourt Assessment. *Journal of Psychoeducational Assessment*, *25*(2), 180–190. https://doi.org/10.1177/0734282906297199

Alexander, B., Kelly, C. E., Adamson, C., Beare, R., Zannino, D., Chen, J., Murray, A. L., Loh, W. Y., Matthews, L. G., Warfield, S. K., Anderson, P. J., Doyle, L. W., Seal, M. L., Spittle, A. J., Cheong, J. L. Y., & Thompson, D. K. (2019). Changes in neonatal regional brain volume associated with preterm birth and perinatal factors. *NeuroImage*, *185*, 654–663. https://doi.org/10.1016/J.NEUROIMAGE.2018.07.021

Allen, M. (2017). *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781483381411

Ambalavanan, N., Carlo, W. A., Tyson, J. E., Langer, J. C., Walsh, M. C., Parikh, N. A., Das, A., Van Meurs, K. P., Shankaran, S., Stoll, B. J., Higgins, R. D., Generic Database, for the G., & Subcommittees of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network, F.-U. S. of the E. K. S. N. I. of C. H. and H. D. N. R. (2012). Outcome trajectories in extremely preterm infants. *Pediatrics*, *130*(1), e115-25. https://doi.org/10.1542/peds.2011-3693

Anblagan, D., Pataky, R., Evans, M. J., Telford, E. J., Serag, A., Sparrow, S., Piyasena, C., Semple, S. I., Wilkinson, A. G., Bastin, M. E., & Boardman, J. P. (2016). Association between preterm brain injury and exposure to chorioamnionitis during fetal life. *Scientific Reports*, *6*, 37932. https://doi.org/10.1038/srep37932

Andescavage, N. N., du Plessis, A., McCarter, R., Serag, A., Evangelou, I., Vezina, G., Robertson, R., & Limperopoulos, C. (2016). Complex Trajectories of Brain Development in the Healthy Human Fetus. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhw306

Anjari, M., Counsell, S. J., Srinivasan, L., Allsop, J. M., Hajnal, J. V., Rutherford, M. A., & Edwards, A. D. (2009). The Association of Lung Disease With Cerebral White Matter Abnormalities in Preterm Infants. *Pediatrics*, *124*(1), 268–276. https://doi.org/10.1542/peds.2008-1294

Armstrong, K. H., & Agazzi, H. C. (2010). The Bayley-III Cognitive Scale. In *Bayley-III Clinical Use and Interpretation* (pp. 29–45). Elsevier. https://doi.org/10.1016/B978-0-12-374177-6.10002-9

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bailey, D. B., Hebbeler, K., Spiker, D., Scarborough, A., Mallik, S., & Nelson, L. (2005). Thirty-six-month outcomes for families of children who have disabilities and participated in early intervention. *Pediatrics*, *116*(6), 1346–1352. https://doi.org/10.1542/peds.2004-1239

Ball, G., Aljabar, P., Nongena, P., Kennea, N., Gonzalez-Cinca, N., Falconer, S., Chew, A. T. M., Harper, N., Wurie, J., Rutherford, M. A., Counsell, S. J., & Edwards, A. D. (2017). Multimodal image analysis of clinical influences on preterm brain development. *Annals of Neurology*, *82*(2), 233–246. https://doi.org/10.1002/ana.24995

Ball, G., Boardman, J. P., Rueckert, D., Aljabar, P., Arichi, T., Merchant, N., Gousias, I. S., Edwards, A. D., & Counsell, S. J. (2012). The effect of preterm birth on thalamic and

cortical development. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(5), 1016–1024. https://doi.org/10.1093/CERCOR/BHR176

Ball, G., Counsell, S. J., Anjari, M., Merchant, N., Arichi, T., Doria, V., Rutherford, M. A., Edwards, A. D., Rueckert, D., & Boardman, J. P. (2010). An optimised tract-based spatial statistics protocol for neonates: applications to prematurity and chronic lung disease. *NeuroImage*, *53*(1), 94–102. https://doi.org/10.1016/J.NEUROIMAGE.2010.05.055

BAPM. (2008). *Report of a BAPM/RCPCH Working Group Classification of health status at 2 years as a perinatal outcome Classification of health status at 2 years as a perinatal outcome*. https://www.networks.nhs.uk/nhs-networks/staffordshire-shropshire-and-black-country-newborn/documents/2_year_Outcome_BAPM_WG_report_v6_Jan08.pdf

Batalle, D., Edwards, A. D., & O'Muircheartaigh, J. (2018). Annual Research Review: Not just a small adult brain: understanding later neurodevelopment through imaging the neonatal brain. In *Journal of Child Psychology and Psychiatry and Allied Disciplines* (Vol. 59, Issue 4, pp. 350–371). Blackwell Publishing Ltd. https://doi.org/10.1111/jcpp.12838

Bauer, P. M., Hanson, J. L., Pierson, R. K., Davidson, R. J., & Pollak, S. D. (2009). Cerebellar Volume and Cognitive Functioning in Children Who Experienced Early Deprivation. *Biological Psychiatry*, *66*(12), 1100–1106. https://doi.org/10.1016/j.biopsych.2009.06.014

Beauport, L., Schneider, J., Faouzi, M., Hagmann, P., Hüppi, P. S., Tolsa, J.-F., Truttmann, A. C., & Fischer Fumeaux, C. J. (2017). Impact of Early Nutritional Intake on Preterm Brain: A Magnetic Resonance Imaging Study. *The Journal of Pediatrics*, *181*, 29-36.e1. https://doi.org/10.1016/j.jpeds.2016.09.073

Belfort, M. B., Anderson, P. J., Nowak, V. A., Lee, K. J., Molesworth, C., Thompson, D. K., Doyle, L. W., & Inder, T. E. (2016). Breast Milk Feeding, Brain Development, and Neurocognitive Outcomes: A 7-Year Longitudinal Study in Infants Born at Less Than 30 Weeks' Gestation. *The Journal of Pediatrics*, *177*, 133-139.e1. https://doi.org/10.1016/j.jpeds.2016.06.045

Belfort, M. B., & Inder, T. E. (2022). Human Milk and Preterm Infant Brain Development: A Narrative Review. *Clinical Therapeutics*, *44*(4), 612–621. https://doi.org/10.1016/j.clinthera.2022.02.011

Bellman, R. (1966). Dynamic programming. *Science (New York, N.Y.)*, *153*(3731), 34–37. https://doi.org/10.1126/science.153.3731.34

Benavente-Fernández, I., Synnes, A., Grunau, R. E., Chau, V., Ramraj, C., Glass, T., Cayam-Rand, D., Siddiqi, A., & Miller, S. P. (2019). Association of Socioeconomic Status and Brain Injury With Neurodevelopmental Outcomes of Very Preterm Children. *JAMA Network Open*, *2*(5), e192914. https://doi.org/10.1001/jamanetworkopen.2019.2914

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). *Pearson Correlation Coefficient* (pp. 1–4). https://doi.org/10.1007/978-3-642-00296-0_5

Betancourt, L. M., Avants, B., Farah, M. J., Brodsky, N. L., Wu, J., Ashtari, M., & Hurt, H. (2016). Effect of socioeconomic status (SES) disparity on neural development in female African-American infants at age 1 month. *Developmental Science*, *19*(6), 947–956. https://doi.org/10.1111/desc.12344

Binder, C., Buchmayer, J., Thajer, A., Giordano, V., Schmidbauer, V., Harreiter, K., Klebermass-Schrehof, K., Berger, A., & Goeral, K. (2021). Association between Fat-Free Mass and Brain Size in Extremely Preterm Infants. *Nutrients*, *13*(12), 4205. https://doi.org/10.3390/nu13124205

Binder, C., Longford, N., Gale, C., Modi, N., & Uthaya, S. (2018). Body Composition following Necrotising Enterocolitis in Preterm Infants. *Neonatology*, *113*(3), 242–248. https://doi.org/10.1159/000485827

Blesa, M., Galdi, P., Sullivan, G., Wheater, E. N., Stoye, D. Q., Lamb, G. J., Quigley, A. J., Thrippleton, M. J., Bastin, M. E., & Boardman, J. P. (2020). Peak Width of Skeletonized Water Diffusion MRI in the Neonatal Brain. *Frontiers in Neurology*, *11*, 235. https://doi.org/10.3389/fneur.2020.00235

Blesa, M., Sullivan, G., Anblagan, D., Telford, E. J., Quigley, A. J., Sparrow, S. A., Serag, A., Semple, S. I., Bastin, M. E., & Boardman, J. P. (2019). Early breast milk exposure modifies brain connectivity in preterm infants. *NeuroImage*, *184*, 431–439. https://www.sciencedirect.com/science/article/pii/S1053811918308309

Boardman, J. P., & Counsell, S. J. (2020). Invited Review: Factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. *Neuropathology and Applied Neurobiology*, *46*(5), 413–421. https://doi.org/10.1111/nan.12589

Boardman, J. P., Counsell, S. J., Rueckert, D., Hajnal, J. V, Bhatia, K. K., Srinivasan, L., Kapellou, O., Aljabar, P., Dyet, L. E., Rutherford, M. A., Allsop, J. M., & Edwards, A. D. (2007). Early growth in brain volume is preserved in the majority of preterm infants. *Annals of Neurology*, *62*(2), 185–192. https://doi.org/10.1002/ANA.21171

Boardman, J. P., Counsell, S. J., Rueckert, D., Kapellou, O., Bhatia, K. K., Aljabar, P., Hajnal, J., Allsop, J. M., Rutherford, M. A., & Edwardsa, A. D. (2006). Abnormal deep grey matter development following preterm birth detected using deformation-based morphometry. *NeuroImage*, *32*(1), 70–78. https://doi.org/10.1016/J.NEUROIMAGE.2006.03.029

Boardman, J. P., Hall, J., Thrippleton, M. J., Reynolds, R. M., Bogaert, D., Davidson, D. J., Schwarze, J., Drake, A. J., Chandran, S., Bastin, M. E., & Fletcher-Watson, S. (2020). Impact of preterm birth on brain development and long-term outcome: protocol for a cohort study in Scotland. *BMJ Open*, *10*(3), e035854. https://doi.org/10.1136/bmjopen-2019-035854

Bode, M. M., D'Eugenio, D. B., Mettelman, B. B., & Gross, S. J. (2014). Predictive Validity of the Bayley, Third Edition at 2 Years for Intelligence Quotient at 4 Years in Preterm Infants. *Journal of Developmental & Behavioral Pediatrics*, *35*(9), 570–575. https://doi.org/10.1097/DBP.0000000000000110

Brady, N. C., Anderson, C. J., Hahn, L. J., Obermeier, S. M., & Kapa, L. L. (2014). Eye tracking as a measure of receptive vocabulary in children with autism spectrum disorders. *Augmentative and Alternative Communication (Baltimore, Md.: 1985)*, *30*(2), 147–159. https://doi.org/10.3109/07434618.2014.904923

Breeman, L. D., Jaekel, J., Baumann, N., Bartmann, P., & Wolke, D. (2015). Preterm Cognitive Function Into Adulthood. *Pediatrics*, *136*(3), 415–423. https://doi.org/10.1542/peds.2015-0608

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breinbauer, C., Mancil, T. L., & Greenspan, S. (2010). The Bayley-III Social-Emotional Scale. *Bayley-III Clinical Use and Interpretation*, 147–175. https://doi.org/10.1016/B978-0-12-374177-6.10005-4

Brossard-Racine, M., Poretti, A., Murnick, J., Bouyssi-Kobar, M., McCarter, R., du Plessis, A. J., & Limperopoulos, C. (2017). Cerebellar Microstructural Organization is Altered by Complications of Premature Birth: A Case-Control Study. *The Journal of Pediatrics*, *182*, 28-33.e1. https://doi.org/10.1016/j.jpeds.2016.10.034

Brouwer, M. J., Kersbergen, K. J., Van Kooij, B. J. M., Benders, M. J. N. L., Van Haastert, I. C., Koopman-Esseboom, C., Neil, J. J., De Vries, L. S., Kidokoro, H., Inder, T. E., & Groenendaal, F. (2017). Preterm brain injury on term-equivalent age MRI in relation to perinatal factors and neurodevelopmental outcome at two years. *PLoS ONE*, *12*(5). https://doi.org/10.1371/JOURNAL.PONE.0177128

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Case-Smith, J., & Alexander, H. (2010). The Bayley-III Motor Scale. *Bayley-III Clinical Use and Interpretation*, 77–146. https://doi.org/10.1016/B978-0-12-374177-6.10004-2

Cavanagh, J., Krishnadas, R., Batty, G. D., Burns, H., Deans, K. A., Ford, I., McConnachie, A., McGinty, A., McLean, J. S., Millar, K., Sattar, N., Shiels, P. G., Tannahill, C., Velupillai, Y. N., Packard, C. J., & McLean, J. (2013). Socioeconomic Status and the Cerebellar Grey Matter Volume. Data from a Well-Characterised Population Sample. *The Cerebellum*, *12*(6), 882–891. https://doi.org/10.1007/s12311-013-0497-4

Chawla, S., Natarajan, G., Shankaran, S., Pappas, A., Stoll, B. J., Carlo, W. A., Saha, S., Das, A., Laptook, A. R., Higgins, R. D., & National Institute of Child Health and Human Development Neonatal Research Network. (2016). Association of Neurodevelopmental Outcomes and Neonatal Morbidities of Extremely Premature Infants With Differential Exposure to Antenatal Steroids. *JAMA Pediatrics*, *170*(12), 1164–1172. https://doi.org/10.1001/jamapediatrics.2016.1936

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Clouchoux, C., Guizard, N., Evans, A. C., du Plessis, A. J., & Limperopoulos, C. (2012). Normative fetal brain growth by quantitative in vivo magnetic resonance imaging. *American Journal of Obstetrics and Gynecology*, *206*(2), 173.e1-173.e8. https://doi.org/10.1016/j.ajog.2011.10.002

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences Third Edition*.

Conti-Ramsden, G., Mok, P. L. H., Pickles, A., & Durkin, K. (2013). Adolescents with a history of specific language impairment (SLI): Strengths and difficulties in social, emotional and behavioral functioning. *Research in Developmental Disabilities*, *34*(11), 4161–4169. https://doi.org/10.1016/J.RIDD.2013.08.043

Cortés-Albornoz, M. C., García-Guáqueta, D. P., Velez-van-Meerbeke, A., & Talero-Gutiérrez, C. (2021). Maternal Nutrition and Neurodevelopment: A Scoping Review. *Nutrients*, *13*(10), 3530. https://doi.org/10.3390/nu13103530

Cost, S., & Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, *10*, 57–78. https://doi.org/https://doi.org/10.1023/A:1022664626993

Counsell, S. J., Arichi, T., Arulkumaran, S., & Rutherford, M. A. (2019). *Fetal and neonatal neuroimaging* (pp. 67–103). https://doi.org/10.1016/B978-0-444-64029-1.00004-7

Counsell, S. J., Edwards, A. D., Chew, A. T. M., Anjari, M., Dyet, L. E., Srinivasan, L., Boardman, J. P., Allsop, J. M., Hajnal, J. V, Rutherford, M. A., & Cowan, F. M. (2008). Specific relations between neurodevelopmental abilities and white matter microstructure in children born preterm. *Brain : A Journal of Neurology*, *131*(Pt 12), 3201–3208. https://doi.org/10.1093/brain/awn268

Coviello, C., Keunen, K., Kersbergen, K. J., Groenendaal, F., Leemans, A., Peels, B., Isgum, I., Viergever, M. A., de Vries, L. S., Buonocore, G., Carnielli, V. P., & Benders, M. J. N. L. (2018). Effects of early nutrition and growth on brain volumes, white matter

microstructure, and neurodevelopmental outcome in preterm newborns. *Pediatric Research*, *83*(1), 102–110. https://doi.org/10.1038/pr.2017.227

Crais, E. R. (2010). The Bayley-III Language Scale. *Bayley-III Clinical Use and Interpretation*, 47–75. https://doi.org/10.1016/B978-0-12-374177-6.10003-0

Crump, C. (2020). An overview of adult health outcomes after preterm birth. *Early Human Development*, *150*, 105187. https://doi.org/10.1016/j.earlhumdev.2020.105187

Davis, E. P., & Sandman, C. A. (2012). Prenatal psychobiological predictors of anxiety risk in preadolescent children. *Psychoneuroendocrinology*, *37*(8), 1224–1233. https://doi.org/10.1016/j.psyneuen.2011.12.016

Dean, D. C., Planalp, E. M., Wooten, W., Kecskemeti, S. R., Adluru, N., Schmidt, C. K., Frye, C., Birn, R. M., Burghy, C. A., Schmidt, N. L., Styner, M. A., Short, S. J., Kalin, N. H., Goldsmith, H. H., Alexander, A. L., & Davidson, R. J. (2018). Association of Prenatal Maternal Depression and Anxiety Symptoms With Infant White Matter Microstructure. *JAMA Pediatrics*, *172*(10), 973. https://doi.org/10.1001/jamapediatrics.2018.2132

Demirci, G. M., Kittler, P. M., Phan, H. T. T., Gordon, A. D., Flory, M. J., Parab, S. M., & Tsai, C.-L. (2024). Predicting mental and psychomotor delay in very pre-term infants using machine learning. *Pediatric Research*, *95*(3), 668–678. https://doi.org/10.1038/s41390-023-02713-z

Dibble, M., Ang, J. Z., Mariga, L., Molloy, E. J., & Bokde, A. L. W. (2021). Diffusion Tensor Imaging in Very Preterm, Moderate-Late Preterm and Term-Born Neonates: A Systematic Review. *The Journal of Pediatrics*, *232*, 48-58.e3. https://doi.org/10.1016/j.jpeds.2021.01.008

Dimitrova, R., Arulkumaran, S., Carney, O., Chew, A., Falconer, S., Ciarrusta, J., Wolfers, T., Batalle, D., Cordero-Grande, L., Price, A. N., Teixeira, R. P. A. G., Hughes, E., Egloff, A., Hutter, J., Makropoulos, A., Robinson, E. C., Schuh, A., Vecchiato, K., Steinweg, J. K., … Edwards, A. D. (2021). Phenotyping the Preterm Brain: Characterizing Individual Deviations From Normative Volumetric Development in Two Large Infant Cohorts. *Cerebral Cortex*, *31*(8), 3665–3677. https://doi.org/10.1093/cercor/bhab039

Dodge, Y. (2008). Spearman Rank Correlation Coefficient. In *The Concise Encyclopedia of Statistics* (pp. 502–505). Springer New York. https://doi.org/10.1007/978-0-387-32833-1_379

Donald, K. A., Hendrikse, C. J., Roos, A., Wedderburn, C. J., Subramoney, S., Ringshaw, J. E., Bradford, L., Hoffman, N., Burd, T., Narr, K. L., Woods, R. P., Zar, H. J., Joshi, S. H., & Stein, D. J. (2024). Prenatal alcohol exposure and white matter microstructural changes across the first 6–7 years of life: A longitudinal diffusion tensor imaging study of a South African birth cohort. *NeuroImage: Clinical*, *41*, 103572. https://doi.org/10.1016/j.nicl.2024.103572

Duerden, E. G., Grunau, R. E., Chau, V., Groenendaal, F., Guo, T., Chakravarty, M. M., Benders, M., Wagenaar, N., Eijsermans, R., Koopman, C., Synnes, A., Vries, L. de, & Miller, S. P. (2020). Association of early skin breaks and neonatal thalamic maturation. *Neurology*, *95*(24), e3420–e3427. https://doi.org/10.1212/WNL.0000000000010953

Ekblad, M., Korkeila, J., & Lehtonen, L. (2015). Smoking during pregnancy affects foetal brain development. *Acta Paediatrica (Oslo, Norway : 1992)*, *104*(1), 12–18. https://doi.org/10.1111/apa.12791

Ene, D., Der, G., Fletcher-Watson, S., O'Carroll, S., MacKenzie, G., Higgins, M., & Boardman, J. P. (2019). Associations of Socioeconomic Deprivation and Preterm Birth With Speech, Language, and Communication Concerns Among Children Aged 27 to

30 Months. *JAMA Network Open*, *2*(9), e1911027. https://doi.org/10.1001/jamanetworkopen.2019.11027

Feldman, H. M., Lee, E. S., Yeatman, J. D., & Yeom, K. W. (2012). Language and reading skills in school-aged children and adolescents born preterm are associated with white matter properties on diffusion tensor imaging. *Neuropsychologia*, *50*(14), 3348–3362. https://doi.org/10.1016/j.neuropsychologia.2012.10.014

Fernández de Gamarra-Oca, L., Ojeda, N., Gómez-Gastiasoro, A., Peña, J., Ibarretxe-Bilbao, N., García-Guerrero, M. A., Loureiro, B., & Zubiaurre-Elorza, L. (2021). Long-Term Neurodevelopmental Outcomes after Moderate and Late Preterm Birth: A Systematic Review. *The Journal of Pediatrics*, *237*, 168-176.e11. https://doi.org/10.1016/j.jpeds.2021.06.004

Finke, E. H., Wilkinson, K. M., & Hickerson, B. D. (2017). Social Referencing Gaze Behavior During a Videogame Task: Eye Tracking Evidence from Children With and Without ASD. *Journal of Autism and Developmental Disorders*, *47*(2), 415–423. https://doi.org/10.1007/s10803-016-2968-1

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gentile, S. (2017). Untreated depression during pregnancy: Short- and long-term effects in offspring. A systematic review. *Neuroscience*, *342*, 154–166. https://doi.org/10.1016/j.neuroscience.2015.09.001

Gillespie-Smith, K., Boardman, J. P., Murray, I. C., Norman, J. E., O'Hare, A., & Fletcher-Watson, S. (2016). Multiple Measures of Fixation on Social Content in Infancy: Evidence for a Single Social Cognitive Construct? *Infancy : The Official Journal of the International Society on Infant Studies*, *21*(2), 241–257. https://doi.org/10.1111/infa.12103

Gilmore, J. H., Lin, W., Prastawa, M. W., Looney, C. B., Vetsa, Y. S. K., Knickmeyer, R. C., Evans, D. D., Smith, J. K., Hamer, R. M., Lieberman, J. A., & Gerig, G. (2007). Regional gray matter growth, sexual dimorphism, and cerebral asymmetry in the neonatal brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(6), 1255–1260. https://doi.org/10.1523/JNEUROSCI.3339-06.2007

Glass, T. J. A., Chau, V., Gardiner, J., Foong, J., Vinall, J., Zwicker, J. G., Grunau, R. E., Synnes, A., Poskitt, K. J., & Miller, S. P. (2017). Severe retinopathy of prematurity predicts delayed white matter maturation and poorer neurodevelopment. *Archives of Disease in Childhood. Fetal and Neonatal Edition*, *102*(6), F532–F537. https://doi.org/10.1136/archdischild-2016-312533

Granger, C., Spittle, A. J., Walsh, J., Pyman, J., Anderson, P. J., Thompson, D. K., Lee, K. J., Coleman, L., Dagia, C., Doyle, L. W., & Cheong, J. (2018). Histologic chorioamnionitis in preterm infants: correlation with brain magnetic resonance imaging at term equivalent age. *BMC Pediatrics*, *18*(1), 63. https://doi.org/10.1186/s12887-018-1001-6

Grizenko, N., Fortier, M.-E., Zadorozny, C., Thakur, G., Schmitz, N., Duval, R., & Joober, R. (2012). Maternal Stress during Pregnancy, ADHD Symptomatology in Children and Genotype: Gene-Environment Interaction. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie Canadienne de Psychiatrie de l'enfant et de l'adolescent*, *21*(1), 9–15.

Guillot, M., Guo, T., Ufkes, S., Schneider, J., Synnes, A., Chau, V., Grunau, R. E., & Miller, S. P. (2020). Mechanical Ventilation Duration, Brainstem Development, and Neurodevelopment in Children Born Preterm: A Prospective Cohort Study. *The Journal of Pediatrics*, *226*, 87-95.e3. https://doi.org/10.1016/J.JPEDS.2020.05.039

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. In *Journal of Machine Learning Research* (Vol. 3).

Hanson, J. L., Chandra, A., Wolfe, B. L., & Pollak, S. D. (2011). Association between Income and the Hippocampus. *PLoS ONE*, *6*(5), e18712. https://doi.org/10.1371/journal.pone.0018712

Hanson, J. L., Hair, N., Shen, D. G., Shi, F., Gilmore, J. H., Wolfe, B. L., & Pollak, S. D. (2013). Family Poverty Affects the Rate of Human Infant Brain Growth. *PLoS ONE*, *8*(12), e80954. https://doi.org/10.1371/journal.pone.0080954

Harman, J. L., & Smith-Bonahue, T. M. (2010). The Bayley-III Adaptive Behavior Scale. *Bayley-III Clinical Use and Interpretation*, 177–200. https://doi.org/10.1016/B978-0-12-374177-6.10006-6

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction* (2nd ed.). Springer New York.

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st ed.). Wiley-IEEE Press.

Herting, M. M., Younan, D., Campbell, C. E., & Chen, J.-C. (2019). Outdoor Air Pollution and Brain Structure and Function From Across Childhood to Young Adulthood: A Methodological Review of Brain MRI Studies. *Frontiers in Public Health*, *7*. https://doi.org/10.3389/fpubh.2019.00332

Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, *2015*, 198363. https://doi.org/10.1155/2015/198363

Hoff Esbjørn, B., Hansen, B. M., Greisen, G., & Mortensen, E. L. (2006). Intellectual development in a Danish cohort of prematurely born preschool children: specific or general difficulties? *Journal of Developmental and Behavioral Pediatrics : JDBP*, *27*(6), 477–484. https://doi.org/10.1097/00004703-200612000-00004

Hua, X., Petrou, S., Coathup, V., Carson, C., Kurinczuk, J. J., Quigley, M. A., Boyle, E., Johnson, S., Macfarlane, A., & Rivero-Arias, O. (2023). Gestational age and hospital admission costs from birth to childhood: a population-based record linkage study in England. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *108*(5), 485–491. https://doi.org/10.1136/archdischild-2022-324763

Inder, T. E., Warfield, S. K., Wang, H., Hüppi, P. S., & Volpe, J. J. (2005). Abnormal Cerebral Structure Is Present at Term in Premature Infants. *Pediatrics*, *115*(2), 286–294. https://doi.org/10.1542/peds.2004-0326

Inder, T. E., Wells, S. J., Mogridge, N. B., Spencer, C., & Volpe, J. J. (2003). Defining the nature of the cerebral abnormalities in the premature infant: a qualitative magnetic resonance imaging study. *The Journal of Pediatrics*, *143*(2), 171–179. https://doi.org/10.1067/S0022-3476(03)00357-3

Jain, V. G., Kline, J. E., He, L., Kline-Fath, B. M., Altaye, M., Muglia, L. J., DeFranco, E. A., Ambalavanan, N., Parikh, N. A., & Cincinnati Infant Neurodevelopment Early Prediction Study Investigators. (2022). Acute histologic chorioamnionitis independently and directly increases the risk for brain abnormalities seen on magnetic resonance imaging in very preterm infants. *American Journal of Obstetrics and Gynecology*, *227*(4), 623.e1-623.e13. https://doi.org/10.1016/j.ajog.2022.05.042

James, G., Witten, D., Hastie, T., & Tiibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*.

Jednoróg, K., Altarelli, I., Monzalvo, K., Fluss, J., Dubois, J., Billard, C., Dehaene-Lambertz, G., & Ramus, F. (2012). The Influence of Socioeconomic Status on Children's Brain Structure. *PLoS ONE*, *7*(8), e42486. https://doi.org/10.1371/journal.pone.0042486

Jha, S. C., Xia, K., Ahn, M., Girault, J. B., Li, G., Wang, L., Shen, D., Zou, F., Zhu, H., Styner, M., Gilmore, J. H., & Knickmeyer, R. C. (2019). Environmental Influences on Infant Cortical Thickness and Surface Area. *Cerebral Cortex*, *29*(3), 1139–1149. https://doi.org/10.1093/CERCOR/BHY020

Johnson, S. B., Riis, J. L., & Noble, K. G. (2016). State of the Art Review: Poverty and the Developing Brain. *Pediatrics*, *137*(4). https://doi.org/10.1542/peds.2015-3075

Johnson, S., & Marlow, N. (2017). Early and long-term outcome of infants born extremely preterm. *Archives of Disease in Childhood*, *102*(1), 97–102. https://doi.org/10.1136/archdischild-2015-309581

Johnson, S., Moore, T., & Marlow, N. (2014). Using the Bayley-III to assess neurodevelopmental delay: which cut-off should be used? *Pediatric Research*, *75*(5), 670–674. https://doi.org/10.1038/pr.2014.10

Kamino, D., Chau, V., Studholme, C., Liu, M., Xu, D., Barkovich, A. J., Ferriero, D. M., Miller, S. P., Brant, R., & Tam, E. W. Y. (2019). Plasma cholesterol levels and brain development in preterm newborns. *Pediatric Research*, *85*(3), 299–304. https://doi.org/10.1038/s41390-018-0260-0

Kapellou, O., Counsell, S. J., Kennea, N., Dyet, L., Saeed, N., Stark, J., Maalouf, E., Duggan, P., Ajayi-Obe, M., Hajnal, J., Allsop, J. M., Boardman, J., Rutherford, M. A., Cowan, F., & Edwards, A. D. (2006). Abnormal Cortical Development after Premature Birth Shown by Altered Allometric Scaling of Brain Growth. *PLoS Medicine*, *3*(8), e265. https://doi.org/10.1371/journal.pmed.0030265

Kaur, S., Powell, S., He, L., Pierson, C. R., & Parikh, N. A. (2014). Reliability and Repeatability of Quantitative Tractography Methods for Mapping Structural White Matter Connectivity in Preterm and Term Infants at Term-Equivalent Age. *PLoS ONE*, *9*(1), e85807. https://doi.org/10.1371/journal.pone.0085807

Kelly, C. E., Thompson, D. K., Cheong, J. L., Chen, J., Olsen, J. E., Eeles, A. L., Walsh, J. M., Seal, M. L., Anderson, P. J., Doyle, L. W., & Spittle, A. J. (2019). Brain structure and neurological and behavioural functioning in infants born preterm. *Developmental Medicine & Child Neurology*, *61*(7), 820–831. https://doi.org/10.1111/dmcn.14084

Kersbergen, K. J., Makropoulos, A., Aljabar, P., Groenendaal, F., de Vries, L. S., Counsell, S. J., & Benders, M. J. N. L. (2016). Longitudinal Regional Brain Development and Clinical Risk Factors in Extremely Preterm Infants. *The Journal of Pediatrics*, *178*, 93-100.e6. https://doi.org/10.1016/j.jpeds.2016.08.024

Khurana, S., Kane, A. E., Brown, S. E., Tarver, T., & Dusing, S. C. (2020). Effect of neonatal therapy on the motor, cognitive, and behavioral development of infants born preterm: a systematic review. *Developmental Medicine and Child Neurology*, *62*(6), 684–692. https://doi.org/10.1111/dmcn.14485

Kidokoro, H., Anderson, P. J., Doyle, L. W., Woodward, L. J., Neil, J. J., & Inder, T. E. (2014). Brain injury and altered brain growth in preterm infants: predictors and prognosis. *Pediatrics*, *134*(2). https://doi.org/10.1542/PEDS.2013-2336

Kim, D.-Y., Park, H.-K., Kim, N.-S., Hwang, S.-J., & Lee, H. J. (2016). Neonatal diffusion tensor brain imaging predicts later motor outcome in preterm neonates with white matter abnormalities. *Italian Journal of Pediatrics*, *42*(1), 104. https://doi.org/10.1186/s13052-016-0309-9

Kira, K., & Rendell, L. A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. *Undefined*.

Knickmeyer, R. C., Xia, K., Lu, Z., Ahn, M., Jha, S. C., Zou, F., Zhu, H., Styner, M., & Gilmore, J. H. (2016). Impact of Demographic and Obstetric Factors on Infant Brain Volumes: A Population Neuroscience Study. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhw331

Knickmeyer, R. C., Xia, K., Lu, Z., Ahn, M., Jha, S. C., Zou, F., Zhu, H., Styner, M., & Gilmore, J. H. (2017). Impact of Demographic and Obstetric Factors on Infant Brain Volumes: A Population Neuroscience Study. *Cerebral Cortex (New York, N.Y. : 1991)*, *27*(12), 5616–5625. https://doi.org/10.1093/CERCOR/BHW331

Knight, M. J., Smith-Collins, A., Newell, S., Denbow, M., & Kauppinen, R. A. (2018). Cerebral White Matter Maturation Patterns in Preterm Infants: An MRI T2 Relaxation Anisotropy and Diffusion Tensor Imaging Study. *Journal of Neuroimaging*, *28*(1), 86–94. https://doi.org/10.1111/jon.12486

Kojima, K., Kline, J. E., Altaye, M., Kline-Fath, B. M., Parikh, N. A., & Cincinnati Infant Neurodevelopment Early Prediction Study (CINEPS) Investigators. (2024). Corpus Callosum Abnormalities at Term-Equivalent Age Are Associated with Language Development at 2 Years' Corrected Age in Infants Born Very Preterm. *Journal of Pediatrics. Clinical Practice*, *11*, 200101. https://doi.org/10.1016/j.jpedcp.2024.200101

Kononenko, I. (1994). *Estimating attributes: Analysis and extensions of RELIEF* (pp. 171–182). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-57868-4_57

Kornbrot, D. (2014). Point Biserial Correlation. In *Wiley StatsRef: Statistics Reference Online*. Wiley. https://doi.org/10.1002/9781118445112.stat06227

Kostović, I., & Jovanov-Milošević, N. (2006). The development of cerebral connections during the first 20–45 weeks' gestation. *Seminars in Fetal and Neonatal Medicine*, *11*(6), 415–422. https://doi.org/10.1016/j.siny.2006.07.001

Kostovic, I., & Vasung, L. (2009). Insights From In Vitro Fetal Magnetic Resonance Imaging of Cerebral Development. *Seminars in Perinatology*, *33*(4), 220–233. https://doi.org/10.1053/j.semperi.2009.04.003

Kraska-MIller, M. (2013). *Nonparametric Statistics for Social and Behavioral Sciences*. Chapman and Hall/CRC. https://doi.org/10.1201/b16188

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Lautarescu, A., Pecheva, D., Nosarti, C., Nihouarn, J., Zhang, H., Victor, S., Craig, M., Edwards, A. D., & Counsell, S. J. (2020). Maternal Prenatal Stress Is Associated With Altered Uncinate Fasciculus Microstructure in Premature Neonates. *Biological Psychiatry*, *87*(6), 559–569. https://doi.org/10.1016/j.biopsych.2019.08.010

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling Developmental Language Difficulties From School Entry Into Adulthood: Literacy, Mental Health, and Employment Outcomes. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1401–1416. https://doi.org/10.1044/1092-4388(2009/08-0142)

Le Dieu-Lugon, B., Dupré, N., Legouez, L., Leroux, P., Gonzalez, B. J., Marret, S., Leroux-Nicollet, I., & Cleren, C. (2020). Why considering sexual differences is necessary when studying encephalopathy of prematurity through rodent models. *European Journal of Neuroscience*, *52*(1), 2560–2574. https://doi.org/10.1111/ejn.14664

Lechner, B. E., & Vohr, B. R. (2017). Neurodevelopmental Outcomes of Preterm Infants Fed Human Milk. *Clinics in Perinatology*, *44*(1), 69–83. https://doi.org/10.1016/j.clp.2016.11.004

Lee, S. H., Lee, S. M., Lim, N. G., Kim, H. J., Bae, S.-H., Ock, M., Kim, U.-N., Lee, J. Y., & Jo, M.-W. (2016). Differences in pregnancy outcomes, prenatal care utilization, and maternal complications between teenagers and adult women in Korea. *Medicine*, *95*(34), e4630. https://doi.org/10.1097/MD.0000000000004630

Limperopoulos, C., Soul, J. S., Gauvreau, K., Huppi, P. S., Warfield, S. K., Bassan, H., Robertson, R. L., Volpe, J. J., & Du Plessis, A. J. (2005). Late gestation cerebellar growth is rapid and impeded by premature birth. *Pediatrics*, *115*(3), 688–695. https://doi.org/10.1542/PEDS.2004-1169

Lind, A., Haataja, L., Rautava, L., Väliaho, A., Lehtonen, L., Lapinleimu, H., Parkkola, R., & Korkman, M. (2010). Relations between brain volumes, neuropsychological assessment and parental questionnaire in prematurely born children. *European Child & Adolescent Psychiatry*, *19*(5), 407–417. https://doi.org/10.1007/s00787-009-0070-3

Lind, A., Parkkola, R., Lehtonen, L., Munck, P., Maunu, J., Lapinleimu, H., & Haataja, L. (2011). Associations between regional brain volumes at term-equivalent age and development at 2 years of age in preterm children. *Pediatric Radiology*, *41*(8), 953–961. https://doi.org/10.1007/S00247-011-2071-X

Linsell, L., Johnson, S., Wolke, D., O'Reilly, H., Morris, J. K., Kurinczuk, J. J., & Marlow, N. (2018). Cognitive trajectories from infancy to early adulthood following birth before 26 weeks of gestation: a prospective, population-based cohort study. *Archives of Disease in Childhood*, *103*(4), 363–370. https://doi.org/10.1136/archdischild-2017-313414

Loh, W. Y., Anderson, P. J., Cheong, J. L. Y., Spittle, A. J., Chen, J., Lee, K. J., Molesworth, C., Inder, T. E., Connelly, A., Doyle, L. W., & Thompson, D. K. (2017). Neonatal basal ganglia and thalamic volumes: Very preterm birth and 7-year neurodevelopmental outcomes. *Pediatric Research*, *82*(6), 970. https://doi.org/10.1038/PR.2017.161

Lu, Y.-C., Kapse, K., Andersen, N., Quistorff, J., Lopez, C., Fry, A., Cheng, J., Andescavage, N., Wu, Y., Espinosa, K., Vezina, G., du Plessis, A., & Limperopoulos, C. (2021). Association Between Socioeconomic Status and In Utero Fetal Brain Development. *JAMA Network Open*, *4*(3), e213526. https://doi.org/10.1001/jamanetworkopen.2021.3526

Luby, J. L., Belden, A. C., Whalen, D., Harms, M. P., & Barch, D. M. (2016). Breastfeeding and Childhood IQ: The Mediating Role of Gray Matter Volume. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(5), 367–375. https://doi.org/10.1016/j.jaac.2016.02.009

MacDonald, M. G., & Seshia, M. M. K. (2015). *Avery's Neonatology: Pathophysiology and Management of the Newborn* (7th ed.). Lippincott Williams and Wilkins.

Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., Edwards, A. D., Counsell, S. J., & Rueckert, D. (2014). Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain. *IEEE Transactions on Medical Imaging*, *33*(9), 1818–1831. https://doi.org/10.1109/TMI.2014.2322280

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E. P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J.-D., … Rueckert, D. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, *173*, 88–112. https://doi.org/10.1016/j.neuroimage.2018.01.054

Mangham, L. J., Petrou, S., Doyle, L. W., Draper, E. S., & Marlow, N. (2009). The Cost of Preterm Birth Throughout Childhood in England and Wales. *Pediatrics*, *123*(2), e312–e327. https://doi.org/10.1542/peds.2008-1827

Markkula, A., Pyhälä-Neuvonen, R., & Stolt, S. (2024). Interventions and their efficacy in supporting language development among preterm children aged 0-3 years – A systematic review. *Early Human Development*, 106057. https://doi.org/10.1016/j.earlhumdev.2024.106057

Matthews, L. G., Inder, T. E., Pascoe, L., Kapur, K., Lee, K. J., Monson, B. B., Doyle, L. W., Thompson, D. K., & Anderson, P. J. (2018). Longitudinal Preterm Cerebellar Volume: Perinatal and Neurodevelopmental Outcome Associations. *The Cerebellum*, *17*(5), 610–627. https://doi.org/10.1007/s12311-018-0946-1

McCarthy, M. M. (2008). Estradiol and the Developing Brain. *Physiological Reviews*, *88*(1), 91–134. https://doi.org/10.1152/physrev.00010.2007

McGoldrick, E., Stewart, F., Parker, R., & Dalziel, S. R. (2020). Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. *Cochrane Database of Systematic Reviews*, *2021*(2). https://doi.org/10.1002/14651858.CD004454.pub4

McManus, B. M., Carle, A. C., & Poehlmann, J. (2012). Effectiveness of part C early intervention physical, occupational, and speech therapy services for preterm or low birth weight infants in Wisconsin, United States. *Academic Pediatrics*, *12*(2), 96–103. https://doi.org/10.1016/j.acap.2011.11.004

Mehler, K., Oberthuer, A., Keller, T., Becker, I., Valter, M., Roth, B., & Kribs, A. (2016). Survival Among Infants Born at 22 or 23 Weeks' Gestation Following Active Prenatal and Postnatal Care. *JAMA Pediatrics*, *170*(7), 671. https://doi.org/10.1001/jamapediatrics.2016.0207

Merz, E. C., Tottenham, N., & Noble, K. G. (2018). Socioeconomic Status, Amygdala Volume, and Internalizing Symptoms in Children and Adolescents. *Journal of Clinical Child & Adolescent Psychology*, *47*(2), 312–323. https://doi.org/10.1080/15374416.2017.1326122

Mewes, A. U. J., Hüppi, P. S., Als, H., Rybicki, F. J., Inder, T. E., McAnulty, G. B., Mulkern, R. V, Robertson, R. L., Rivkin, M. J., & Warfield, S. K. (2006). Regional Brain Development in Serial Magnetic Resonance Imaging of Low-Risk Preterm Infants. *Pediatrics*, *118*(1), 23–33. https://doi.org/10.1542/PEDS.2005-2675

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *The American Psychologist*, *56*(2), 128–165.

Miller, J., Tonkin, E., Damarell, R., McPhee, A., Suganuma, M., Suganuma, H., Middleton, P., Makrides, M., & Collins, C. (2018). A Systematic Review and Meta-Analysis of Human Milk Feeding and Morbidity in Very Low Birth Weight Infants. *Nutrients*, *10*(6), 707. https://doi.org/10.3390/nu10060707

Miller, S. L., Huppi, P. S., & Mallard, C. (2016). The consequences of fetal growth restriction on brain structure and neurodevelopmental outcome. *The Journal of Physiology*, *594*(4), 807–823. https://doi.org/10.1113/JP271402

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (Eds.). (2014). *Handbook of Missing Data Methodology*. Chapman and Hall/CRC. https://doi.org/10.1201/b17622

Murray, A. L., Thompson, D. K., Pascoe, L., Leemans, A., Inder, T. E., Doyle, L. W., Anderson, J. F. I., & Anderson, P. J. (2016). White matter abnormalities and impaired attention abilities in children born very preterm. *NeuroImage*, *124*(Pt A), 75–84. https://doi.org/10.1016/j.neuroimage.2015.08.044

Myrhaug, H. T., Brurberg, K. G., Hov, L., & Markestad, T. (2019). Survival and Impairment of Extremely Premature Infants: A Meta-analysis. *Pediatrics*, *143*(2). https://doi.org/10.1542/peds.2018-0933

National Guideline Alliance (UK). (2017). *Developmental follow-up of children and young people born preterm*. National Institute for Health and Care Excellence (NICE).

Nguyen The Tich, S., Anderson, P. J., Hunt, R. W., Lee, K. J., Doyle, L. W., & Inder, T. E. (2011). Neurodevelopmental and perinatal correlates of simple brain metrics in very preterm infants. *Archives of Pediatrics & Adolescent Medicine*, *165*(3), 216–222. https://doi.org/10.1001/ARCHPEDIATRICS.2011.9

Nishida, M., Makris, N., Kennedy, D. N., Vangel, M., Fischl, B., Krishnamoorthy, K. S., Caviness, V. S., & Grant, P. E. (2006). Detailed semiautomated MRI based

morphometry of the neonatal brain: Preliminary results. *NeuroImage*, *32*(3), 1041–1049. https://doi.org/10.1016/j.neuroimage.2006.05.020

Norman, M., Hallberg, B., Abrahamsson, T., Björklund, L. J., Domellöf, M., Farooqi, A., Foyn Bruun, C., Gadsbøll, C., Hellström-Westas, L., Ingemansson, F., Källén, K., Ley, D., Maršál, K., Normann, E., Serenius, F., Stephansson, O., Stigson, L., Um-Bergström, P., & Håkansson, S. (2019). Association Between Year of Birth and 1-Year Survival Among Extremely Preterm Infants in Sweden During 2004-2007 and 2014-2016. *JAMA*, *321*(12), 1188. https://doi.org/10.1001/jama.2019.2021

Oakland, T. (2011). Adaptive Behavior Assessment System – Second Edition. In *Encyclopedia of Clinical Neuropsychology* (pp. 37–39). Springer New York. https://doi.org/10.1007/978-0-387-79948-3_1506

Ohuma, E. O., Moller, A.-B., Bradley, E., Chakwera, S., Hussain-Alkhateeb, L., Lewin, A., Okwaraji, Y. B., Mahanani, W. R., Johansson, E. W., Lavin, T., Fernandez, D. E., Domínguez, G. G., de Costa, A., Cresswell, J. A., Krasevec, J., Lawn, J. E., Blencowe, H., Requejo, J., & Moran, A. C. (2023). National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *The Lancet*, *402*(10409), 1261–1271. https://doi.org/10.1016/S0140-6736(23)00878-4

Ortiz-Mantilla, S., Choudhury, N., Leevers, H., & Benasich, A. A. (2008). Understanding language and cognitive deficits in very low birth weight children. *Developmental Psychobiology*, *50*(2), 107–126. https://doi.org/10.1002/dev.20278

Orton, J., Doyle, L. W., Tripathi, T., Boyd, R., Anderson, P. J., & Spittle, A. (2024). Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database of Systematic Reviews*, *2024*(2). https://doi.org/10.1002/14651858.CD005495.pub5

Ottolini, K. M., Andescavage, N., Kapse, K., Jacobs, M., & Limperopoulos, C. (2020). Improved brain growth and microstructural development in breast milk–fed very low birth weight premature infants. *Acta Paediatrica*, *109*(8), 1580–1587. https://doi.org/10.1111/apa.15168

Owen, D., & Matthews, S. G. (2003). Glucocorticoids and Sex-Dependent Development of Brain Glucocorticoid and Mineralocorticoid Receptors. *Endocrinology*, *144*(7), 2775–2784. https://doi.org/10.1210/en.2002-0145

Pannek, K., Fripp, J., George, J. M., Fiori, S., Colditz, P. B., Boyd, R. N., & Rose, S. E. (2018). Fixel-based analysis reveals alterations is brain microstructure and macrostructure of preterm-born infants at term equivalent age. *NeuroImage: Clinical*, *18*, 51–59. https://doi.org/10.1016/j.nicl.2018.01.003

Paredes, I., Hidalgo, L., Chedraui, P., Palma, J., & Eugenio, J. (2005). Factors associated with inadequate prenatal care in Ecuadorian women. *International Journal of Gynecology & Obstetrics*, *88*(2), 168–172. https://doi.org/10.1016/j.ijgo.2004.09.024

Parikh, N. A., Lasky, R. E., Kennedy, K. A., McDavid, G., & Tyson, J. E. (2013). Perinatal Factors and Regional Brain Volume Abnormalities at Term in a Cohort of Extremely Low Birth Weight Infants. *PLoS ONE*, *8*(5), e62804. https://doi.org/10.1371/journal.pone.0062804

Parker, M. G., Stellwagen, L. M., Noble, L., Kim, J. H., Poindexter, B. B., & Puopolo, K. M. (2021). Promoting Human Milk and Breastfeeding for the Very Low Birth Weight Infant. *Pediatrics*, *148*(5). https://doi.org/10.1542/peds.2021-054272

Partridge, S. C., Mukherjee, P., Henry, R. G., Miller, S. P., Berman, J. I., Jin, H., Lu, Y., Glenn, O. A., Ferriero, D. M., Barkovich, A. J., & Vigneron, D. B. (2004). Diffusion tensor imaging: serial quantitation of white matter tract maturity in premature newborns. *NeuroImage*, *22*(3), 1302–1314. https://doi.org/10.1016/j.neuroimage.2004.02.038

Pascal, A., Govaert, P., Oostra, A., Naulaers, G., Ortibus, E., & Van den Broeck, C. (2018). Neurodevelopmental outcome in very preterm and very-low-birthweight infants born over the past decade: a meta-analytic review. *Developmental Medicine & Child Neurology*, *60*(4), 342–355. https://doi.org/10.1111/dmcn.13675

Patel, R. M., Rysavy, M. A., Bell, E. F., & Tyson, J. E. (2017). Survival of Infants Born at Periviable Gestational Ages. *Clinics in Perinatology*, *44*(2), 287–303. https://doi.org/10.1016/j.clp.2017.01.009

Pecheva, D., Kelly, C., Kimpton, J., Bonthrone, A., Batalle, D., Zhang, H., & Counsell, S. J. (2018). Recent advances in diffusion neuroimaging: applications in the developing preterm brain. *F1000Research*, *7*. https://doi.org/10.12688/f1000research.15073.1

Pecheva, D., Tournier, J.-D., Pietsch, M., Christiaens, D., Batalle, D., Alexander, D. C., Hajnal, J. V., Edwards, A. D., Zhang, H., & Counsell, S. J. (2019). Fixel-based analysis of the preterm brain: Disentangling bundle-specific white matter microstructural and macrostructural changes in relation to clinical risk factors. *NeuroImage: Clinical*, *23*, 101820. https://doi.org/10.1016/j.nicl.2019.101820

Perin, J., Mulick, A., Yeung, D., Villavicencio, F., Lopez, G., Strong, K. L., Prieto-Merino, D., Cousens, S., Black, R. E., & Liu, L. (2022). Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet Child & Adolescent Health*, *6*(2), 106–115. https://doi.org/10.1016/S2352-4642(21)00311-4

Peterson, B. S., Anderson, A. W., Ehrenkranz, R., Staib, L. H., Tageldin, M., Colson, E., Gore, J. C., Duncan, C. C., Makuch, R., & Ment, L. R. (2003). Regional brain volumes and their later neurodevelopmental correlates in term and preterm infants. *Pediatrics*, *111*(5 Pt 1), 939–948. https://doi.org/10.1542/PEDS.111.5.939

Piñon, M. (2010). Theoretical Background and Structure of the Bayley Scales of Infant and Toddler Development, Third Edition. *Bayley-III Clinical Use and Interpretation*, 1–28. https://doi.org/10.1016/B978-0-12-374177-6.10001-7

Pogribna, U., Yu, X., Burson, K., Zhou, Y., Lasky, R. E., Narayana, P. A., & Parikh, N. A. (2013). Perinatal Clinical Antecedents of White Matter Microstructural Abnormalities on Diffusion Tensor Imaging in Extremely Preterm Infants. *PLoS ONE*, *8*(8), e72974. https://doi.org/10.1371/journal.pone.0072974

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions of the Institute of British Geographers*, *52*, 145. https://doi.org/10.2307/621706

Power, V. A., Spittle, A. J., Lee, K. J., Anderson, P. J., Thompson, D. K., Doyle, L. W., & Cheong, J. L. Y. (2019). Nutrition, Growth, Brain Volume, and Neurodevelopment in Very Preterm Children. *The Journal of Pediatrics*, *215*, 50-55.e3. https://doi.org/10.1016/j.jpeds.2019.08.031

Raghunathan, T. E., Lepkowski, J., Hoewyk, J. Van, & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Undefined*.

Rajagopalan, V., Scott, J., Habas, P. A., Kim, K., Corbett-Detig, J., Rousseau, F., Barkovich, A. J., Glenn, O. A., & Studholme, C. (2011). Local tissue growth patterns underlying normal fetal human brain gyrification quantified in utero. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(8), 2878–2887. https://doi.org/10.1523/JNEUROSCI.5458-10.2011

Ramel, S. E., & Georgieff, M. K. (2014). *Preterm Nutrition and the Brain* (pp. 190–200). https://doi.org/10.1159/000358467

Reiss, A. L., Kesler, S. R., Vohr, B., Duncan, C. C., Katz, K. H., Pajot, S., Schneider, K. C., Makuch, R. W., & Ment, L. R. (2004). Sex differences in cerebral volumes of 8-year-

olds born preterm. *The Journal of Pediatrics*, *145*(2), 242–249. https://doi.org/10.1016/J.JPEDS.2004.04.031

Rifkin-Graboi, A., Meaney, M. J., Chen, H., Bai, J., Hameed, W. B., Tint, M. T., Broekman, B. F. P., Chong, Y.-S., Gluckman, P. D., Fortier, M. V., & Qiu, A. (2015). Antenatal Maternal Anxiety Predicts Variations in Neural Structures Implicated in Anxiety Disorders in Newborns. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(4), 313-321.e2. https://doi.org/10.1016/j.jaac.2015.01.013

Robinson, M., Mattes, E., Oddy, W. H., Pennell, C. E., van Eekelen, A., McLean, N. J., Jacoby, P., Li, J., De Klerk, N. H., Zubrick, S. R., Stanley, F. J., & Newnham, J. P. (2011). Prenatal stress and risk of behavioral morbidity from age 2 to 14 years: The influence of the number, type, and timing of stressful life events. *Development and Psychopathology*, *23*(2), 507–520. https://doi.org/10.1017/S0954579411000241

Robnik-Sikonja, M., & Kononenko, I. (1997). *An adaptation of Relief for attribute estimation in regression*.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, *53*(1/2), 23–69. https://doi.org/10.1023/A:1025667309714

Rogers, C. E., Smyser, T., Smyser, C. D., Shimony, J., Inder, T. E., & Neil, J. J. (2016). Regional white matter development in very preterm infants: perinatal predictors and early developmental outcomes. *Pediatric Research*, *79*(1), 87–95. https://doi.org/10.1038/pr.2015.172

Rose, J., Cahill-Rowley, K., Vassar, R., Yeom, K. W., Stecher, X., Stevenson, D. K., Hintz, S. R., & Barnea-Goraly, N. (2015). Neonatal brain microstructure correlates of neurodevelopment and gait in preterm children 18-22 mo of age: an MRI and DTI study. *Pediatric Research*, *78*(6), 700–708. https://doi.org/10.1038/pr.2015.157

Rose, S. E., Hatzigeorgiou, X., Strudwick, M. W., Durbridge, G., Davies, P. S. W., & Colditz, P. B. (2008). Altered white matter diffusion anisotropy in normal and preterm infants at term-equivalent age. *Magnetic Resonance in Medicine*, *60*(4), 761–767. https://doi.org/10.1002/mrm.21689

Ross, A., & Willson, V. L. (2017). Paired Samples T-Test. In *Basic and Advanced Statistical Tests* (pp. 17–19). SensePublishers. https://doi.org/10.1007/978-94-6351-086-8_4

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Ruigrok, A. N. V, Salimi-Khorshidi, G., Lai, M. C., Baron-Cohen, S., Lombardo, M. V, Tait, R. J., & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, *39*, 34–50. https://doi.org/10.1016/J.NEUBIOREV.2013.12.004

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sarda, S. P., Sarri, G., & Siffel, C. (2021). Global prevalence of long-term neurodevelopmental impairment following extremely preterm birth: a systematic literature review. *Journal of International Medical Research*, *49*(7), 030006052110280. https://doi.org/10.1177/03000605211028026

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International*, *107*(44), 776–782. https://doi.org/10.3238/arztebl.2010.0776

Schneider, J., Fischer Fumeaux, C. J., Duerden, E. G., Guo, T., Foong, J., Graz, M. B., Hagmann, P., Chakravarty, M. M., Hüppi, P. S., Beauport, L., Truttmann, A. C., & Miller,

S. P. (2018). Nutrient Intake in the First Two Weeks of Life and Brain Growth in Preterm Neonates. *Pediatrics*, *141*(3). https://doi.org/10.1542/peds.2017-2169

Scottish National Statistics. (2016). *SIMD - Scottish Index of Multiple Deprivation: SIMD16 Technical Notes*. Scottish National Statistics.

Setänen, S., Lehtonen, L., Parkkola, R., Aho, K., Haataja, L., Ahtola, A., Ekblad, M., Ekblad, S., Ekholm, E., Eurola, A., Huhtala, M., Kero, P., Kiiski-Mäki, H., Korja, R., Lahti, K., Lapinleimu, H., Lehtonen, T., Leppänen, M., Lind, A., … Ylijoki, M. (2016). Prediction of neuromotor outcome in infants born preterm at 11 years of age using volumetric neonatal magnetic resonance imaging and neurological examinations. *Developmental Medicine and Child Neurology*, *58*(7), 721–727. https://doi.org/10.1111/DMCN.13030

Shah, D. K., Doyle, L. W., Anderson, P. J., Bear, M., Daley, A. J., Hunt, R. W., & Inder, T. E. (2008). Adverse neurodevelopment in preterm infants with postnatal sepsis or necrotizing enterocolitis is mediated by white matter abnormalities on magnetic resonance imaging at term. *The Journal of Pediatrics*, *153*(2), 170–175, 175.e1. https://doi.org/10.1016/j.jpeds.2008.02.033

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Spittle, A., Orton, J., Anderson, P. J., Boyd, R., & Doyle, L. W. (2015). Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database of Systematic Reviews*, *11*, CD005495. https://doi.org/10.1002/14651858.CD005495.pub4

Srinivasan, L., Dutta, R., Counsell, S. J., Allsop, J. M., Boardman, J. P., Rutherford, M. A., & Edwards, A. D. (2007). Quantification of deep gray matter in preterm infants at term-equivalent age using manual volumetry of 3-tesla magnetic resonance images. *Pediatrics*, *119*(4), 759–765. https://doi.org/10.1542/PEDS.2006-2508

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Stiver, M. L., Kamino, D., Guo, T., Thompson, A., Duerden, E. G., Taylor, M. J., & Tam, E. W. Y. (2015). Maternal Postsecondary Education Associated With Improved Cerebellar Growth After Preterm Birth. *Journal of Child Neurology*, *30*(12), 1633–1639. https://doi.org/10.1177/0883073815576790

Stoll, B. J., Hansen, N. I., Bell, E. F., Walsh, M. C., Carlo, W. A., Shankaran, S., Laptook, A. R., Sánchez, P. J., Van Meurs, K. P., Wyckoff, M., Das, A., Hale, E. C., Ball, M. B., Newman, N. S., Schibler, K., Poindexter, B. B., Kennedy, K. A., Cotten, C. M., Watterberg, K. L., … Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. (2015). Trends in Care Practices, Morbidity, and Mortality of Extremely Preterm Neonates, 1993-2012. *JAMA*, *314*(10), 1039–1051. https://doi.org/10.1001/jama.2015.10244

Stoye, D. Q., Blesa, M., Sullivan, G., Galdi, P., Lamb, G. J., Black, G. S., Quigley, A. J., Thrippleton, M. J., Bastin, M. E., Reynolds, R. M., & Boardman, J. P. (2020). Maternal cortisol is associated with neonatal amygdala microstructure and connectivity in a sexually dimorphic manner. *ELife*, *9*. https://doi.org/10.7554/eLife.60729

Sullivan, G., Vaher, K., Blesa, M., Galdi, P., Stoye, D. Q., Quigley, A. J., Thrippleton, M. J., Norrie, J., Bastin, M. E., & Boardman, J. P. (2022). Breast Milk Exposure is Associated With Cortical Maturation in Preterm Infants. *Annals of Neurology*. https://doi.org/10.1002/ana.26559

Sveinsdóttir, K., Ley, D., Hövel, H., Fellman, V., Hüppi, P. S., Smith, L. E. H., Hellström, A., & Hansen Pupp, I. (2018). Relation of Retinopathy of Prematurity to Brain Volumes at

Term Equivalent Age and Developmental Outcome at 2 Years of Corrected Age in Very Preterm Infants. *Neonatology*, *114*(1), 46–52. https://doi.org/10.1159/000487847

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Allyn & Bacon/Pearson Education.

Taylor, P. A., Jacobson, S. W., van der Kouwe, A., Molteno, C. D., Chen, G., Wintermark, P., Alhamud, A., Jacobson, J. L., & Meintjes, E. M. (2015). A DTI-based tractography study of effects on brain structure associated with prenatal alcohol exposure in newborns. *Human Brain Mapping*, *36*(1), 170–186. https://doi.org/10.1002/hbm.22620

Telford, E. J., Fletcher-Watson, S., Gillespie-Smith, K., Pataky, R., Sparrow, S., Murray, I. C., O'Hare, A., & Boardman, J. P. (2016). Preterm birth is associated with atypical social orienting in infancy detected using eye tracking. *Journal of Child Psychology and Psychiatry*, *57*(7), 861–868. https://doi.org/10.1111/jcpp.12546

Thompson, D. K., Inder, T. E., Faggian, N., Johnston, L., Warfield, S. K., Anderson, P. J., Doyle, L. W., & Egan, G. F. (2011). Characterization of the corpus callosum in very preterm and full-term infants utilizing MRI. *NeuroImage*, *55*(2), 479–490. https://doi.org/10.1016/j.neuroimage.2010.12.025

Thompson, D. K., Kelly, C. E., Chen, J., Beare, R., Alexander, B., Seal, M. L., Lee, K. J., Matthews, L. G., Anderson, P. J., Doyle, L. W., Cheong, J. L. Y., & Spittle, A. J. (2019a). Characterisation of brain volume and microstructure at term-equivalent age in infants born across the gestational age spectrum. *NeuroImage: Clinical*, *21*, 101630. https://doi.org/10.1016/j.nicl.2018.101630

Thompson, D. K., Kelly, C. E., Chen, J., Beare, R., Alexander, B., Seal, M. L., Lee, K., Matthews, L. G., Anderson, P. J., Doyle, L. W., Spittle, A. J., & Cheong, J. L. Y. (2019b). Early life predictors of brain development at term-equivalent age in infants born across the gestational age spectrum. *NeuroImage*, *185*, 813–824. https://doi.org/10.1016/J.NEUROIMAGE.2018.04.031

Thompson, D. K., Warfield, S. K., Carlin, J. B., Pavlovic, M., Wang, H. X., Bear, M., Kean, M. J., Doyle, L. W., Egan, G. F., & Inder, T. E. (2007). Perinatal risk factors altering regional brain structure in the preterm infant. *Brain : A Journal of Neurology*, *130*(Pt 3), 667–677. https://doi.org/10.1093/BRAIN/AWL277

Thompson, D. K., Wood, S. J., Doyle, L. W., Warfield, S. K., Lodygensky, G. A., Anderson, P. J., Egan, G. F., & Inder, T. E. (2008). Neonate hippocampal volumes: prematurity, perinatal predictors, and 2-year outcome. *Annals of Neurology*, *63*(5), 642–651. https://doi.org/10.1002/ANA.21367

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Torras-Mañá, M., Guillamón-Valenzuela, M., Ramírez-Mallafré, A., Brun-Gasca, C., & Fornieles-Deu, A. (2014). Usefulness of the Bayley scales of infant and toddler development,third edition, in the early diagnosis of language disorder. *Psicothema*, *26*(3), 349–356. https://doi.org/10.7334/psicothema2014.29

Triplett, R. L., Lean, R. E., Parikh, A., Miller, J. P., Alexopoulos, D., Kaplan, S., Meyer, D., Adamson, C., Smyser, T. A., Rogers, C. E., Barch, D. M., Warner, B., Luby, J. L., & Smyser, C. D. (2022). Association of Prenatal Exposure to Early-Life Adversity With Neonatal Brain Volumes at Birth. *JAMA Network Open*, *5*(4), e227045. https://doi.org/10.1001/jamanetworkopen.2022.7045

Trittmann, J. K., Nelin, L. D., & Klebanoff, M. A. (2013). Bronchopulmonary dysplasia and neurodevelopmental outcome in extremely preterm neonates. *European Journal of Pediatrics*, *172*(9), 1173–1180. https://doi.org/10.1007/s00431-013-2016-5

Tsanas, A. (2022). Relevance, redundancy, and complementarity trade-off (RRCT): A principled, generic, robust feature-selection tool. *Patterns*, *3*(5), 100471. https://doi.org/10.1016/j.patter.2022.100471

Tsanas, A., Little, M. A., & McSharry, P. E. (2013). A methodology for the analysis of medical data. In *Handbook of Systems and Complexity in Health* (pp. 113–125). Springer.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, *59*(5), 1264–1271. https://doi.org/10.1109/TBME.2012.2183367

Tyson, J. E., Parikh, N. A., Langer, J., Green, C., Higgins, R. D., & National Institute of Child Health and Human Development Neonatal Research Network. (2008). Intensive care for extreme prematurity--moving beyond gestational age. *The New England Journal of Medicine*, *358*(16), 1672–1681. https://doi.org/10.1056/NEJMoa073059

Valavani, E., Blesa, M., Galdi, P., Sullivan, G., Dean, B., Cruickshank, H., Sitko-Rudnicka, M., Bastin, M. E., Chin, R. F. M., MacIntyre, D. J., Fletcher-Watson, S., Boardman, J. P., & Tsanas, A. (2021). Language function following preterm birth: prediction using machine learning. *Pediatric Research*. https://doi.org/10.1038/s41390-021-01779-x

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463

van Noort-van der Spek, I. L., Franken, M.-C. J. P., & Weisglas-Kuperus, N. (2012). Language functions in preterm-born children: a systematic review and meta-analysis. *Pediatrics*, *129*(4), 745–754. https://doi.org/10.1542/peds.2011-1728

Vandormael, C., Schoenhals, L., Hüppi, P. S., Filippa, M., & Borradori Tolsa, C. (2019). Language in Preterm Born Children: Atypical Development and Effects of Early Interventions on Neuroplasticity. *Neural Plasticity*, *2019*, 6873270. https://doi.org/10.1155/2019/6873270

Vassar, R., Schadl, K., Cahill-Rowley, K., Yeom, K., Stevenson, D., & Rose, J. (2020). Neonatal Brain Microstructure and Machine-Learning-Based Prediction of Early Language Development in Children Born Very Preterm. *Pediatric Neurology*. https://doi.org/10.1016/J.PEDIATRNEUROL.2020.02.007

Vesoulis, Z. A., El Ters, N. M., Herco, M., Whitehead, H. V, & Mathur, A. M. (2018). A Web-Based Calculator for the Prediction of Severe Neurodevelopmental Impairment in Preterm Infants Using Clinical and Imaging Characteristics. *Children (Basel, Switzerland)*, *5*(11). https://doi.org/10.3390/children5110151

Villar, J., Giuliani, F., Fenton, T. R., Ohuma, E. O., Ismail, L. C., & Kennedy, S. H. (2016). INTERGROWTH-21st very preterm size at birth reference charts. *The Lancet*, *387*(10021), 844–845. https://doi.org/10.1016/S0140-6736(16)00384-6

Villar, J., Ismail, L. C., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D. G., Lambert, A., Papageorghiou, A. T., Carvalho, M., Jaffer, Y. A., Gravett, M. G., Purwar, M., Frederick, I. O., Noble, A. J., Pang, R., Barros, F. C., Chumlea, C., Bhutta, Z. A., Kennedy, S. H., & International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). (2014). International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *The Lancet*, *384*(9946), 857–868. https://doi.org/10.1016/S0140-6736(14)60932-6

Volpe, J. J. (2009). Cerebellum of the Premature Infant: Rapidly Developing, Vulnerable, Clinically Important. *Journal of Child Neurology*, *24*(9), 1085–1104. https://doi.org/10.1177/0883073809338067

Volpe, J. J. (2018). *Volpe's Neurology of the Newborn*. Elsevier. https://doi.org/10.1016/C2010-0-68825-0

Volpe, J. J. (2019). Dysmaturation of Premature Brain: Importance, Cellular Mechanisms, and Potential Interventions. *Pediatric Neurology*, *95*, 42–66. https://doi.org/10.1016/j.pediatrneurol.2019.02.016

Wang, N., Cui, L., Liu, Z., Wang, Y., Zhang, Y., Shi, C., & Cheng, Y. (2021). Optimizing parenteral nutrition to achieve an adequate weight gain according to the current guidelines in preterm infants with birth weight less than 1500 g: a prospective observational study. *BMC Pediatrics*, *21*(1), 303. https://doi.org/10.1186/s12887-021-02782-1

WHO. (2018). *Preterm birth*. https://www.who.int/en/news-room/fact-sheets/detail/preterm-birth

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

Wolke, D., Samara, M., Bracewell, M., Marlow, N., & EPICure Study Group. (2008). Specific language difficulties and school achievement in children born at 25 weeks of gestation or less. *The Journal of Pediatrics*, *152*(2), 256–262. https://doi.org/10.1016/j.jpeds.2007.06.043

Woodward, L. J., Edgin, J. O., Thompson, D., & Inder, T. E. (2005). Object working memory deficits predicted by early brain injury and development in the preterm infant. *Brain*, *128*(11), 2578–2587. https://doi.org/10.1093/brain/awh618

World Health Organization. (2015). *WHO recommendations on Interventions to Improve Preterm Birth Outcomes*.

Young, J. M., Powell, T. L., Morgan, B. R., Card, D., Lee, W., Smith, M. Lou, Sled, J. G., & Taylor, M. J. (2015). Deep grey matter growth predicts neurodevelopmental outcomes in very preterm children. *NeuroImage*, *111*, 360–368. https://doi.org/10.1016/J.NEUROIMAGE.2015.02.030

Zwicker, J. G., Miller, S. P., Grunau, R. E., Chau, V., Brant, R., Studholme, C., Liu, M., Synnes, A., Poskitt, K. J., Stiver, M. L., & Tam, E. W. Y. (2016). Smaller Cerebellar Growth and Poorer Neurodevelopmental Outcomes in Very Preterm Infants Exposed to Neonatal Morphine. *The Journal of Pediatrics*, *172*, 81-87.e2. https://doi.org/10.1016/j.jpeds.2015.12.024