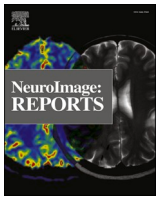
Neuroimage: Reports 3 (2023) 100170

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports



Predicting neurodevelopmental outcomes from neonatal cortical microstructure: A conceptual replication study

Andrea Gondová [a,b,*], Sara Neumane [a,b,d], Yann Leprince [b], Jean-François Mangin [c], Tomoki Arichi [d,e], Jessica Dubois [a,b]

[a] Université Paris Cité, Inserm, NeuroDiderot, F-75019, Paris, France
[b] Université Paris-Saclay, CEA, NeuroSpin UNIACT, F-91191, Gif-sur-Yvette, France
[c] Université Paris-Saclay, CEA, CNRS, NeuroSpin, BAOBAB, Gif-sur-Yvette, France
[d] Centre for the Developing Brain, School of Biomedical Engineering and Imaging Sciences, King's College London, London, SE1 7EH, United Kingdom
[e] Paediatric Neurosciences, Evelina London Children's Hospital, Guy's and St Thomas' NHS Foundation Trust, London, SE1 7EH, United Kingdom

ARTICLE INFO

Keywords:
Brain development
Neonates
Prematurity
DTI (Diffusion tensor imaging)
ML (machine learning)
Prediction
Generalisability

ABSTRACT

Machine learning combined with large-scale neuroimaging databases has been proposed as a promising tool for improving our understanding of the behavioural emergence and early prediction of the neurodevelopmental outcome. A recent example of this strategy is a study by Ouyang et al. (2020) which suggested that cortical microstructure quantified by diffusion MRI through fractional anisotropy (FA) metric in preterm and full-term neonates can lead to effective prediction of language and cognitive outcomes at 2 years of corrected age as assessed by *Bayley Scales of Infant and Toddler Development, Third Edition* (BSID-III) composite scores. Given the important need for robust and generalisable tools which can reliably predict the neurodevelopmental outcome of preterm infants, we aimed to replicate the conclusions of this work using a larger independent dataset from the *developing Human Connectome Project* dataset (dHCP, third release) with early MRI and BSID-III evaluation at 18 months of corrected age. We then aimed to extend the validation of the proposed predictive pipeline through the study of different cohorts (the largest one included 295 neonates, with gestational age between 29 and 42 week and post-menstrual age at MRI between 31 and 45 weeks). This allowed us to evaluate whether some limitations of the original study (mainly small sample size and limited variability in the input and output features used in the predictive models) would influence the prediction results. In contrast to the original study that inspired the current work, our prediction results did not outcompete the random levels. Furthermore, these negative results persisted even when the study settings were expanded. Our findings suggest that the cortical microstructure close to birth described by DTI-FA measures might not be sufficient for a reliable prediction of BSID-III scores during toddlerhood, at least in the current setting, i.e. generally older cohorts and a different processing pipeline. Our inability to conceptually replicate the results of the original study is in line with the previously reported replicability issues within the machine learning field and demonstrates the challenges in defining the good set of practices for the implementation and validation of reliable predictive tools in the neurodevelopmental (and other) fields.

1. Introduction

The human brain undergoes dynamic and complex structural and functional development during the pre- and perinatal period (Silbereis et al., 2016; Parikh, 2016; Ouyang et al., 2019a; Kostović et al., 2019), driven by several cellular and molecular mechanisms that show heterogenous progression at both spatial and temporal levels. Deviations from the normal developmental sequence due to events like preterm birth can have profound and long-term effects, including motor, cognitive, and language deficits (Johnston, 2009; Hadaya and Nosarti, 2020). However, these are typically not observed until their acquisition during early childhood (e.g. around 18 months for independent walking) leading to relatively late diagnosis despite the perinatal onset of underlying brain abnormalities (Arpi and Ferrari, 2013; Parikh, 2016; Marín, 2016). As personalised neuroprotective and rehabilitation interventions are likely to be most effective early in infancy when neuroplasticity is enhanced

* Corresponding author. Université Paris Cité, Inserm, NeuroDiderot, F-75019, Paris, France.
E-mail address: andrea.gondova@cea.fr (A. Gondová).

https://doi.org/10.1016/j.ynirp.2023.100170
Received 28 October 2022; Received in revised form 10 March 2023; Accepted 1 April 2023
Available online 14 April 2023

(Blauw-Hospers et al., 2007; Pickler et al., 2010; Müller et al., 2017), the ability to identify at-risk neonates in the first few weeks after birth could have major public health benefits through improving the quality of life of potentially affected children and their families. In this context, markers of abnormal early brain development such as those provided by non-invasive magnetic resonance imaging (MRI) have a clear role in predicting brain development trajectories and infant behavioural outcomes. Moreover, this can be further enhanced by combining extracted features with sophisticated, infant-adapted machine learning (ML) strategies (see Baker and Kandasamy (2022) for a review focused on premature infants).

In this vein, Ouyang et al. (2020) recently proposed one such strategy based on diffusion tensor imaging (DTI) and cortical microstructure at birth using regional fractional anisotropy (FA) measures as inputs. FA is thought to directly reflect the cellular and molecular processes underlying cortical development during the preterm period (e.g. growth of radial glia projections and apical dendrites, development of multi-orientation intra-cortical connections, proliferation of membranes) (Ball et al., 2013; Ouyang et al., 2019a). Ouyang's study thus tested the hypothesis that regional FA measures in preterm and full-term neonates could serve as a useful biomarker for the prediction of distinctive aspects of later neurodevelopmental outcome assessed with *Bayley Scales of Infant and Toddler Development, Third Edition* (BSID-III) at around 2 years of corrected age. Their results suggested that inter-individual variability of cortical microstructure at birth provided sufficient information for successful and robust prediction of later cognitive and language outcomes using linear support vector regression (SVR). Additionally, based on the evaluation of model feature importance, the authors suggested that the distinctive features were encoded in uniquely distributed patterns across the cerebral cortex.

Despite these promising results, the authors of this original study remained cautious given the limitations of their strategy, highlighting the small dataset (46 infants) as one of the major problems affecting the potential generalisability of their proposed method to the wider population. Additionally, motor outcomes in their dataset could not be predicted, although they suggested that it was unclear whether these null results resulted from insufficient heterogeneity within the cohort (in terms of cortical inputs and/or outcome scores) or other factors such as low signal-to-noise ratio (SNR) of cortical FA in primary sensorimotor cortical regions associated with motor function. Given these limitations and the important need for robust and generalisable tools which can reliably predict the neurodevelopmental outcome of preterm infants, this original study seemed to us an ideal target for conceptual replication. As opposed to direct replication, conceptual replication aims to test the generality of the original suggestion, i.e. that cortical FA close to birth is predictive of later neurodevelopmental outcome, with a different but similarly reliable method (Zwaan et al., 2017).

In short, in the present study, we aimed to examine whether cortical microstructure close to birth –assessed with regional FA measures - is predictive of later neurodevelopmental outcomes in an independent dataset with early MRI acquired in neonates and behavioural follow-up as part of the "developing Human Connectome Project" (dHCP). To test this, our goal was to replicate the original study of Ouyang et al. (2020) as closely as possible given the available information. Importantly, we opted for a different but robust processing pipeline: differences are specified in detail within the methods and discussion sections. Additionally, we expanded the predictive pipeline by incorporating model tuning and nested validation for a more robust evaluation of the results. To validate the implementation of the pipeline, we also developed predictive baselines on simpler predictive tasks, i.e. prediction of the gestational age (GA) at birth, infant prematurity status (categorised GA at birth with 37 weeks threshold), as well as categorised behavioural scores.

2. Materials and methods

2.1. Data

We included a sample of preterm and full-term neonates participating in the open-source *developing Human Connectome Project* (http://www.developingconnectome.org/). Detailed subject demographic and clinical information can be found in the section *Cohort Description*. Data collection for the dHCP took place in London, UK from 2015 to 2020 with UK NHS research ethics committee approval (14/LO/1169, IRAS 138070), and written informed consent obtained from parents.

2.2. Cohort Description

By taking advantage of the large sample size of available dHCP data, we were able to perform this replication study on different cohort profiles, starting from the cohort with the largest heterogeneity and subsequently limiting the inclusion criteria to approximate the composition of the original study of Ouyang et al. (2020).

2.2.1. Cohort A

Cohort A included all subjects available in the dHCP 3rd data release with complete diffusion and anatomical MRI which passed the quality control (QC) (Edwards et al., 2022), and with available BSID-III composite scores assessed between 17 and 20 months of age. In cases of subjects with multiple sessions, we prioritised the session at the youngest age to approximate the original study that aimed to scan subjects as close to birth as possible. Additionally, we included only infants that were considered healthy and without major brain focal lesions (i.e. who did not present any visible abnormality of possible clinical significance on structural MRI evaluated by an expert Paediatric Neuroradiologist, recorded in dHCP databases as radiological score range 1–3). To approximate the exclusion criteria described in the original study from the clinical information available in the dHCP database as closely as possible, we additionally excluded 63 additional subjects who had a history of hypoxic-ischemic encephalopathy, lung disease or bronchopulmonary dysplasia, necrotizing enterocolitis requiring intestinal resection or complex feeding/nutritional disorders, maternal drug or alcohol abuse or smoking during pregnancy in their records. Additionally, 12 subjects were excluded as the quality of the registration to the neonate atlas and subsequent projection to the cortex was insufficient (see section *Extraction of cortical diffusion metrics*).

The final Cohort A then included 295 infants (54% males, gestational age at birth –GA at birth: median 39.9 weeks, range [29.9w −42.3w]; scanned at median post-menstrual age –PMA: 40.9w, range [31.1w −45.1w]).

2.2.2. Cohort B

We additionally restricted the subject space by removing infants whose PMA at scan and GA at birth were outside the range of the original study (i.e. [31.9w-41.7w] PMA at scan and [25.0w-41.4w] GA at birth). We thus aimed to investigate whether prediction models can leverage wider age variability or whether the original results were limited to a narrow range of GA at birth and PMA at scan. Additionally, this allowed us to test whether low variability in motor outcome score in the original study may have been the reason for a lack of prediction by leveraging the larger size of the dHCP dataset through selecting for increased variability in the neurodevelopmental outcomes. With the GA at birth and PMA at scan matched as close as possible to the data in the original study, Cohort B then included 196 subjects (57% male, median GA at birth 39.1w, range [29.9w-41.3w], median PMA at scan 39.9, range [32.3w-41.6w]).

2.2.3. Cohort C

Additionally, Cohort C's subjects were further selected from Cohort B based on matching the ranges of BSID-III composite scores ([65–110] for cognitive, [59–112] for language, and [73–107] for motor scores) to the original study. In this setting, the cohort was therefore highly similar to the original one in terms of PMA at scan, GA at birth, and the BSID-III score ranges but had a larger sample size of 126 subjects which could allow investigation of the effect of the training sample size on predictive performance. The subjects within Cohort C were 58% male with median GA at birth 39.1w, range [29.9w-41.3w] and median PMA at scan 40.0, range

[32.7w – 41.6w]) in contrast to the original study which included 72% of males.

2.2.4. Cohort D

Finally, Cohort D served as a close replication of the original study where subjects were limited to the same number (n = 46) with similar ranges of PMA at scan, GA at birth and BSID-III score ranges. Randomly sampling 46 subjects from the Cohort C for the training and testing allowed us to test the robustness of the predictive pipeline relatively to the training set size, and thus the generalisability of the predictions.

The detailed description of the Cohorts A, B, and C can be found in Table 1.

2.3. Description of available data

2.3.1. Diffusion MRI

MRI data was acquired using a Philips 3-T Achieva scanner (Philips Medical Systems, Best, The Netherlands). All infants were scanned during natural sleep in a scanner environment optimized for neonatal imaging including a dedicated transport system, positioning device and an optimally sized neonatal 32-channel receive coil with a custom-made acoustic hood as previously described (Hughes et al., 2017).

The diffusion MRI data was used in its pre-processed state available from the dHCP 3rd data release. In sum, diffusion-weighted (DW) images were acquired following a multi-shell high angular resolution diffusion imaging (HARDI) protocol (with b = 0, 400, 1000, 2600 s/mm2) (Tournier et al., 2019) and pre-processed with correction for motion artefacts and slice-to-volume reconstruction using the SHARD approach, leading to a final isotropic voxel size of 1.5mm (Christiaens et al., 2021; Hutter et al., 2018). All included data passed SHARD QC provided in the dHCP (Edwards et al., 2022). Quantitative metric maps resulting from the diffusion tensor imaging (DTI) model, including fractional anisotropy (FA) maps, were computed based on b = 0 and 1000s/mm2 images using FSL's DTIFIT. For distribution of motion across the dataset as well as example FA maps for two subjects, see Supp. Fig. 1.

2.3.2. Structural MRI

The structural data was acquired and reconstructed following optimized protocols (Cordero-Grande et al., 2018) leading to super-resolved T2w images with an isotropic spatial resolution of 0.5mm. Subsequent processing followed a dedicated pipeline for segmentation and cortical surface extraction for T2w neonatal brain images (Bovzek et al., 2018; Makropoulos et al., 2018) with bias-correction, brain extraction, and segmentation using the Draw-EM (Developing brain Region Annotation with Expectation Maximisation) algorithm (Makropoulos et al., 2014). Available meshes of inner cortical surface (corresponding to white matter surface) were used for the cortical parcellations and extraction of DTI metrics such as FA.

2.3.3. Neurodevelopmental assessment and infant characteristics at 18 months

Neurodevelopmental outcome was assessed at St Thomas' Hospital, London by two experienced assessors (a paediatrician and a chartered psychologist) using the Bayley Scales of Infant and Toddler Development, Third Edition (BSID-III) (Bayley, 2012). We only considered evaluations performed at around 18 months of age (between 17 and 20m; corrected for GA at birth in PT infants). Three distinct developmental categories – motor, cognition, and language functions – were assessed yielding age-standardized respective composite scores, with higher values indicating better infant development and scores below 85 (i.e., lower than –1SD of the mean at 100) indicating a developmental delay.

BSID-III composite scores in the Cohort A were the following: median cognitive score of 100, range [60–130]; median language score of 97, range [47–153]; and median motor score of 103, range [70–127]. The

Table 1
Description of Cohort A, B, C (Cohort D is a random subset of C). BSID-III - Bayley Scales of Infant and toddler Development, 3rd edition; CA – corrected age; mode of delivery: V – vaginal, C_{em} – emergency Caesarean section, C_{el} – elective Caesarean section; PMA – post-menstrual age; std – standard deviation; subjects per template age: categorization of infants to age-standardized template according to their PMA at scan.

	Cohort A (n = 295)		Cohort B (n = 196)		Cohort C (n = 126)	
	Mean (std)	Median [range]	Mean (std)	Median [range]	Mean (std)	Median [range]
PMA at scan (weeks)	40.43 (2.446)	40.86 [31.14, 45.14]	39.29 (2.014)	39.86 [32.29, 41.57]	39.46 (1.793)	40.00 [32.71, 41.57]
GA at birth (weeks)	39.15 (2.361)	39.86 [29.86, 42.29]	38.50 (2.417)	39.14 [29.86, 41.29]	38.65 (2.172)	39.14 [29.86, 41.29]
* Pearson correlation	r = 0.81, p < 0.001*		r = 0.91, p < 0.001*		r = 0.86, p < 0.001*	
Scan – birth delay (weeks)	1.28 (1.469)	0.57 [0.00,8.7]	0.79 (1.033)	0.29 [0.00,6.85]	0.81 (1.120)	0.29 [0.00,6.85]
Age at BSID-III assessment (months CA)	18.17 (0.860)	18.00 [17.00, 20.00]	18.31 (0.827)	18.00 [17.00, 20.00]	18.24 (0.821)	18.00 [17.00, 20.00]
Weight at birth (kg)	3.18 (0.689)	3.30 [0.76, 4.61]	3.03 (0.700)	3.11 [0.76, 4.59]	3.09 (0.646)	3.18 [1.25, 4.59]
Male: number (%)	158 (53.6%)		111 (56.6%)		73 (58.0%)	
White: number (%)	158 (53.6%)		93 (47.5%)		52 (41.3%)	
Mode of delivery: number	V: 155, I: 67, Cem: 73, Cel: 0		V: 102, I: 40, Cem: 54, Cel: 0		V: 60, I: 32, Cem: 34, Cel: 0	
Subject per template age: number	33w: 7, 36w: 29, 39w: 259		33: 6, 36: 29, 39: 161		33: 3, 36: 15, 39: 108	
BSID-III composite scores	**Mean (std)**	**Median [range]**	**Mean (std)**	**Median [range]**	**Mean (std)**	**Median [range]**
Cognitive	101.1 (11.18)	100 [60, 130]	99.6 (11.40)	100 [60, 130]	95.7 (9.14)	95 [65, 110]
Language	97.0 (16.05)	97 [47, 153]	95.4 (16.22)	97 [47, 153]	90.9 (12.81)	91 [59, 112]
Motor	101.7 (9.48)	103 [70, 127]	100.8 (9.78)	100 [70, 127]	97.2 (7.98)	98 [73, 107]

*Pearson correlation between GA at birth and PMA at scan.

score medians and ranges were identical for the Cohort B. In case of the Cohort C, the median BSID-III composite scores were the following: cognitive score of 95, range [60–110]; language score of 91, range [59–112]; and motor score of 98, range [73–107]. Cohort D was a random subset of the Cohort C.

Additionally, family socio-economic status (SES) was measured using the Index of multiple deprivation (IMD) which is a UK geographically defined composite social risk score comprising data on income, employment, health, education, living environment, and crime calculated at 18mCA assessment from the mother's home address at the time of birth.

2.4. Data processing

Additionally, we used cortical surfaces extracted as meshes from the structural MRI data, to obtain a cortical parcellation and extract the DTI metrics in the cortical ribbon.

2.4.1. Extraction of cortical DTI metrics

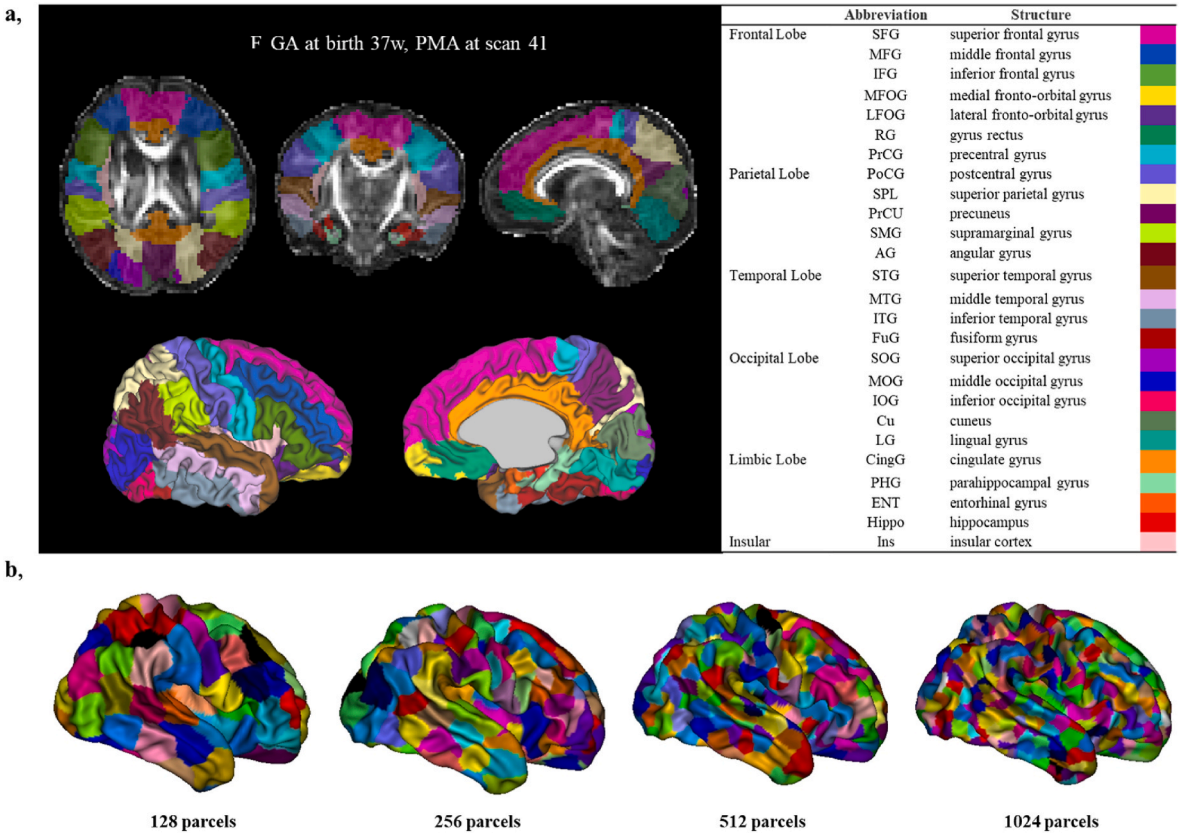
We registered and projected the FA values to the inner cortical surface using a cylindrical approach guided by the spatial location of the local minimum of axial diffusivity (AD). This has been shown to reliably locate the DTI metrics in the cortical ribbon (details available in (Lebenberg et al., 2019)): this method robustly excludes values from surrounding cerebrospinal fluid (CSF) and underlying white matter at the individual level, since these tissues show higher AD values than the grey matter in the immature brain of neonates. This approach for cortical FA extraction has been reliably used and validated in different studies (Lebenberg et al., 2019; Rolland et al., 2019; Adibpour et al., 2020) but it differed from the one of the original study (pipeline not available at the time of this conceptual replication study) which used a cortical skeleton and a fast-marching correction. Visual and quantitative validation suggested 12 subjects with small local mistakes in the metric projection, leading to their exclusion from all cohorts (see section *Cohorts description*).

2.4.2. Cortical parcellation

2.4.2.1. Neonate atlas. As in the original study, FA maps were registered to the neonate segmentation atlas described in Feng et al. (2019) and Oishi et al. (2011). But, as recommended by the atlas' authors (brainmrimap.org, n.d), we found it more reliable to use FSL's FNIRT for the subject-to-atlas projections rather than the large deformation diffeomorphic metric mapping (LDDMM) registration that was used in the original study.

Firstly, subjects were split into 3 groups based on their PMA at scan: the 33w group (31.1, 34.5): 14 subjects; the 36w group [34.5, 37.5): 42 subjects; the 39w group [37.5, 45.1): 239 subjects. The linear transformations were then estimated between subject FA map and the corresponding age-specific atlas spaces, which were then used to initiate the non-linear registration. The resulting warps were inverted and used to project the atlas parcellations to the native DW subject space. Labelling information and the list of the 52 cortical ROIs (26 per hemisphere) used in the study as well as examples of cortical parcellation are presented in Fig. 1a. Segmented labels were then projected to the subject' cortical surface with the same method as the diffusion metric maps.

Visual QC of the registration results and parcellation projections on the cortical mesh were performed with a specific focus on the subjects with the highest distance of their PMA at scan to the atlas template age. We noticed small errors stemming from the registration especially in the cingulate area and regions where the opposite gyri banks were close in volumetric space, which led to multiple parcels per label. To correct the errors, we developed a post-processing step which iteratively reassigned



	Abbreviation	Structure
Frontal Lobe	SFG	superior frontal gyrus
	MFG	middle frontal gyrus
	IFG	inferior frontal gyrus
	MFOG	medial fronto-orbital gyrus
	LFOG	lateral fronto-orbital gyrus
	RG	gyrus rectus
	PrCG	precentral gyrus
Parietal Lobe	PoCG	postcentral gyrus
	SPL	superior parietal gyrus
	PrCU	precuneus
	SMG	supramarginal gyrus
	AG	angular gyrus
Temporal Lobe	STG	superior temporal gyrus
	MTG	middle temporal gyrus
	ITG	inferior temporal gyrus
	FuG	fusiform gyrus
Occipital Lobe	SOG	superior occipital gyrus
	MOG	middle occipital gyrus
	IOG	inferior occipital gyrus
	Cu	cuneus
	LG	lingual gyrus
Limbic Lobe	CingG	cingulate gyrus
	PHG	parahippocampal gyrus
	ENT	entorhinal gyrus
	Hippo	hippocampus
Insular	Ins	insular cortex

Fig. 1. a, Example parcellation of 52 cortical parcels in volumetric view and on the cortical surface. More details on the used atlas can be found in Feng et al. (2019). **b,** Example of random parcellation in the dHCP template space for the different number of parcels.

the small unconnected labelled regions to their nearest neighbour main label, based on the path distance along the cortical surface; this was run until only one parcel per label remained.

2.4.2.2. Random parcellation. Additionally, cortical surfaces were parcellated into random bilateral parcels with similar area (k-means based on Dijkstra-path, optimized to balance the vertex numbers per region). The numbers were more than double the atlas parcels: 128, 256, 512, 1024. The random parcellation was performed in the dHCP surface template space after manual segmentation of the cingulate area, which was excluded from the final parcellations to avoid possible inclusion of non-cortical tissues (Fig. 1b). The random parcellations were then projected to the subjects' anatomical space using the transformations available from the dHCP database.

2.4.2.3. Extraction of DTI metrics in cortical parcels. For both the atlas and the random parcellations, individual DTI metrics were reliably extracted from every cortical region using the DTI maps projected to the subjects' inner cortical surfaces and extracting the median FA per region (as this has similar values to the mean but higher robustness to potential outliers). We then created individual subject feature vectors, considering all cortical parcels and to serve as an input to the subsequent predictive pipeline.

2.4.2.4. Age correction. To address the likely confounding effect of various PMA at scan across subjects on the predictive performance observed during the piloting stages of our replication (see *Results* section), we decided to adjust the regional FA values for PMA at scan. Following on from previous studies, we assumed a relationship between global FA (median over the whole cortical surface) and PMA at scan with an inflection point between 35 and 43w (Batalle et al., 2019; Ouyang et al., 2020). We fitted a continuous piecewise linear regression to Cohort A assuming two linear segments with external knots being the minimal and maximal PMA at scan within the cohort and the middle inflection point between the above specified range. Results of the fits evaluated by Akaike information criterion (AIC), and Bayesian information criterion (BIC) metrics suggested that the ideal inflection point in our data is 36w PMA which was then used as a separation age to perform the corrections over the younger and older age ranges. The line segments were estimated only from the training data and applied to the testing data within the predictive pipeline to avoid data leakage. Although we acknowledge that this inflection point might not be optimal for all parcels given their heterogeneous maturation with age, a single inflection point was preferred for all parcels to avoid heterogeneous FA corrections.

2.5. Predictive pipeline

Wherever possible, the replication pipeline followed the original study as closely as possible given the provided description of methods in the original and referenced publications (Ouyang et al., 2019b, 2020; Yu et al., 2016).

2.5.1. Prediction of continuous composite scores

We recreated the individualised prediction using the SVR algorithm (nu-SVR with a linear kernel and C of 9) to predict continuous variables – cognitive, language, and motor composite BSID-III scores. As the authors of the original study did not report any tuning of the hyperparameters during their implementation or between tasks, we re-used their hyperparameter settings. Given the different sample sizes of our cohorts, we adopted k-fold cross-validation with 46 folds to recreate the training-testing split proportions used in the original study. Leave-one-out cross-validation (LOOCV) is a special case of k-fold where n of the testing set is 1, and thus Cohort D was evaluated exactly like in the original study.

The individual feature vectors of FA metrics from cortical parcels (either from atlas or random parcellation) were then used to predict the outcome scores. During training, features were independently scaled between 0 and 1 using the training splits to determine the scaling factors which were then applied to the testing set. Predictive results were evaluated using a Pearson correlation coefficient (*r)* and mean absolute error (MAE) between the predicted and real composite scores. Additionally, we evaluated the coefficient of determination (R2), which is bounded between (-∞,1] with 0 suggesting the random prediction. In contrast to Pearson's r which is insensitive to scale with potential high correlations possible even for large differences between predicted and actual target value, R2 quantifies a portion of variation in the target predicted by the model and is thus a more indicative measure of predictive performance. However, as the outcome variance might differ between samples, R2 does not allow comparisons across different datasets: MAE measuring the predictive error in the units of the original target measure is then a useful metric.

Replicating the original study, we also performed a permutation test to assess the predictive performance. In short, we randomly shuffled the BSID-III scores 1000 times to generate null distributions for random predictors for both the Pearson's correlation coefficient and MAE. P-values of observing the reported r (or MAE) by chance is then calculated as the ratio of number of permutation tests with r higher than the observed r value over the number of permutation tests (reverse for the MAE).

Additionally, to test the impact of random small differences in the dataset on the robustness of the predictive results, we randomised the inputs and outputs 100 times and repeated the same evaluation to derive the mean and standard deviation of predictive results across the independent runs to determine how sensitive the results are to small changes in the inputs upon shuffling.

We performed the same training and evaluation as a continuous prediction of GA at birth to further validate the implementation of the pipeline.

2.5.2. Categorical predictors

Additionally, we also simplified the predictive task by categorising the independent variables. In the case of BSID-III scores, infants were categorised based on a threshold into Typical (>85) and Atypical (≤85) outcome groups. We then tested prediction of whether the infant developed typically or not, given the cortical FA features using the support vector classifier (SVC) algorithm with predictive results evaluated using Area Under the (Receiver Operating Characteristic) Curve (AUC), Specificity (SPEC), and Sensitivity (SENS) scores. The same permutation testing (reported p-value being the proportion of permutation tests with AUC above the reported AUC divided by number of permutation tests) and shuffling as described in the previous section were also performed.

Similarly, in the case of GA at birth prediction, we performed a classification using categorised GA at birth into *prematurity status* as outputs: preterm (PT) born before 37 weeks of gestation vs full-term (FT), which served to validate the predictive pipeline's implementation as well as the input features.

2.5.3. Hyperparameter tuning & nested validation (Cohort D)

The larger size of the dHCP available data compared to the original work allows us to perform a more thorough validation of the predictive pipeline (Fig. 2). To do this, we randomly selected a subset *n* times (n = 10) of 46 infants from the Cohort C keeping them as an independent testing set (to recreate the validation from the original study). The remaining data, i.e. 80 subjects, was used within the inner validation LOOCV loop (79 subjects to train and one as validation) which allowed us to tune the hyperparameters of the SVR. The best model was then re-trained on the 80 subjects and evaluated on the independent testing split. The mean R2 and MEA results across the 10 repetitions are reported as the final predictive results. Again, the prediction of GA at birth

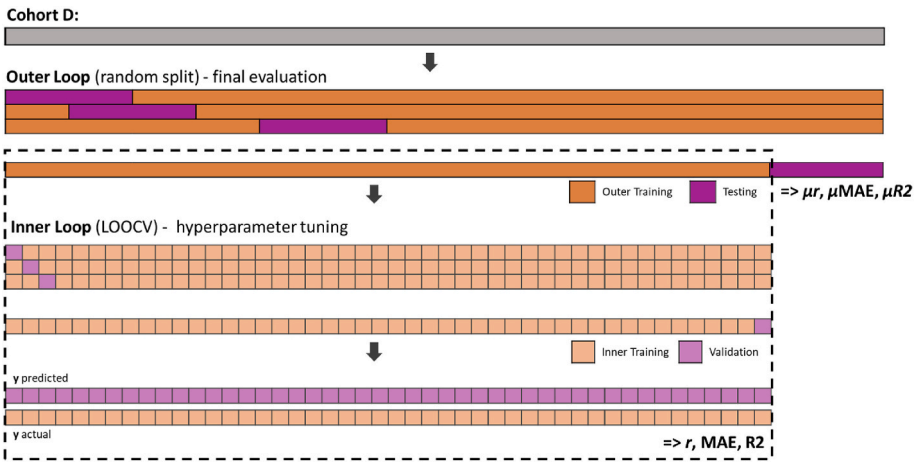


Fig. 2. Nested validation pipeline. μr: mean Pearson's r; μMAE: mean absolute error, μR2: mean R2, n: number of outer loop repetitions, k: number of inner loop repetitions.

and the binarized outcome scores were used as the *sanity* check and performed the same evaluation in the form of a categorical predictive task.

3. Results

3.1. Evaluation of cortical microstructure

As expected, the change in global cortical FA with PMA at scan did not appear to follow a simple linear relationship but showed a steep decrease for young ages followed by an inflection point identified at 36w PMA for Cohort A (Supp. Fig. 1). Additionally, the FA metrics differed significantly between the preterm- and term-born subjects even after correcting for PMA at scan in all cohorts (Supp. Table 1). We also incorporated two major risk factors – GA at birth and sex – in an ANCOVA model to investigate their effect on global cortical FA (corrected for PMA at scan). The results suggested a significant effect of sex on cortical FA, but not of GA at birth, which is likely due to the skew within the Cohort A towards the older neonates.

Visual assessment of the distribution of median FA metrics over the 52 parcels of the atlas for Cohort A (Fig. 3) revealed individual and regional variability in cortical microstructure across the sample. However, it is interesting to note that we observed globally lower FA values in comparison to the original study: more than effects related to PMA at scan, image quality or to the method for extracting cortical FA measures, we suspect that this relates to the SHARD pre-processing pipeline (Christiaens et al., 2021) for correction of motion artefacts and geometric distortions.

3.2. Neurodevelopmental outcomes at 18 months

For Cohort A the BSID-III composite scores at 18 months ranged from 60 to 130 (mean ± std 101.1 ± 11.2) for cognition, 47 to 153 (97.0 ± 16.0) for language, and 70 to 127 (101.7 ± 9.5) for motor outcomes. BSID-III score distribution was very similar in Cohort B. Cohort C showed, by design, more restricted BSID-III scores. The descriptive values are detailed in Table 1. No significant differences between preterm and full-term infants were found in any of the BSID-III scores across cohorts (Fig. 4). Additionally, we did not observe any significant correlation between any specific age (i.e. GA at birth, PMA at scan, age at BSID-III assessment) and neurodevelopmental outcome scores across the cohorts or within preterm and full-term subgroups (Supp. Table 2) except for Cognitive and Language scores and PMA at scan within the FT group of Cohort A after the FDR corrections for multiple comparisons.

Composition of the evaluated cohorts in terms of their Atypical/Typical subject numbers is shown in Table 2. Percentage of PT (or FT) group of the whole PT (or FT) population within the cohort is also shown. The distribution of the PT and FT infants across the Atypical and Typical categories for the 3 scores are very similar, although in the case of motor score in all cohorts, the PT subjects seem to be somewhat overrepresented. Interestingly, in the case of language scores, it is the FT population that seems to have a higher proportion of subjects with Atypical scores.

Additionally, we performed an ANCOVA modelling to study the effect of environment approximated by the IMD on BSID-III scores, and including the global cortical FA and risk factors – GA at birth and sex – within the descriptive analysis (Supp. Table 3). We observed the effects of global cortical FA on Cognitive, Language, but not the Motor scores which is in line with the predictive results presented in the original paper. As expected, IMD (and sex for cognitive and language scores) seemed to affect the neurodevelopmental outcomes.

3.3. Predictive pipeline results

3.3.1. Prematurity status & GA at birth prediction

Prediction of prematurity status (preterm vs full-term categorised depending on GA at birth) using different inputs served two goals: i, validation of the predictive pipeline implementation, and ii, validation of the input features. Overall, cortical FA seemed predictive of prematurity status at birth in all cohorts (Supp. Materials: Predictive results, Fig. 5a). However, the predictive power tended to decrease (sometimes reaching random predictor levels) after the correction of the inputs for PMA at scan which might be explained by the high correlation between PMA at scan and GA at birth in all cohorts (Table 1). This suggests that some, but not all of the prediction was driven by the distribution of this confounding variable, thus stressing the importance of correcting for PMA at scan in all of the following predictive settings. Therefore, all the following reported results are based on the inputs corrected for the PMA at scan.

Interestingly, the random parcellations led to consistently higher predictive results than the atlas parcellation with a tendency of the predictive power to increase with number of regions. This is possibly due to larger number of parcels providing more granular information than when collapsed too much within the atlas parcels.

Another observed tendency was the general trend of decreasing predictive power across the cohorts as the dataset gets progressively limited in variability of inputs and sample sizes. Interestingly, in the case of Cohort D, which is the closest replication of the original study, we observed a substantial drop in predictive power of prematurity status. Comparing these results to those in the Cohort C suggests this is largely

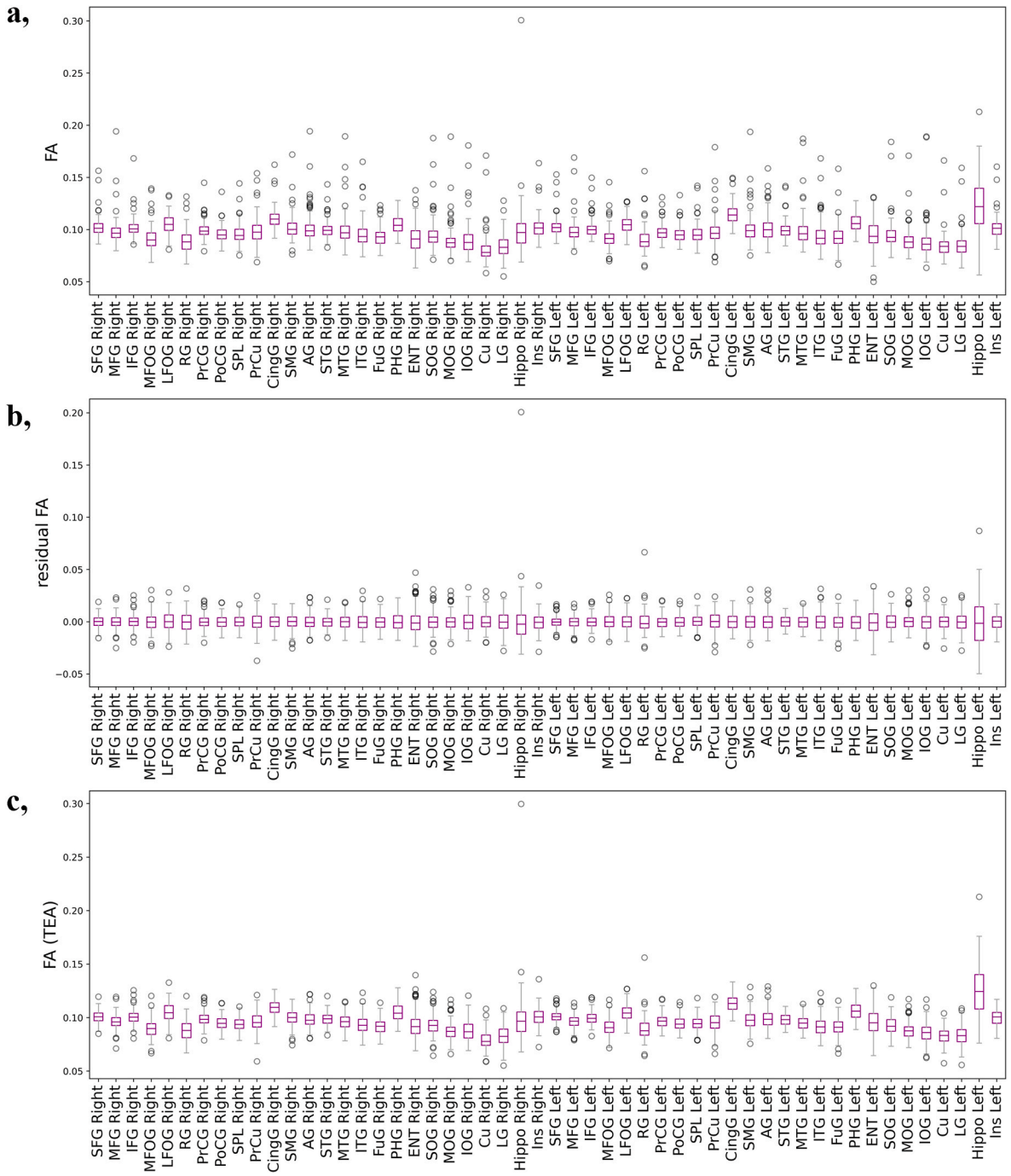


Fig. 3. a, Distribution of median FA per cortical region (atlas parcellation) suggesting inter-individual and inter-regional microstructural variability (Cohort A). **b,** Distribution of the residual FA estimated as the residual of a piecewise linear regression considering the regional FA as the function of PMA at scan with inflection at 36w. **c,** Distribution of the corrected FA after the reintroduction of expected regional FA values at term equivalent age (TEA = 40w PMA) (predicted from the correction models represented in b). Labelling abbreviations are detailed in Fig. 1.

due to insufficient sample size leading to negative results rather than the limited variability within the data given the similar distribution of input and output features.

We also performed a similar prediction aiming to categorise infants into three classes by additionally subdividing the preterm group of Cohort A into extreme-very preterm (EVP, GA at birth <32w, N = 8) and moderate-late preterm (MLP, GA at birth ≥32w, N = 35) groups. In this setting, only the atlas parcellation led to higher than random prediction (Supp. Table 4). Results are in line with previous findings demonstrating the differential effect of the degree of prematurity on brain measures and suggest that the categorization within the preterm group is possible. Nevertheless, it is important to note that practical issues, such as size of the dataset or class imbalance, are likely to impact the reliability of these observations, at least in our limited sample.

When attempting to predict continuous GA at birth rather than categorised prematurity status, the predictive models did not outperform random levels (Supp. Materials: Predictive results, Fig. 5c). Overall, these results at the same time validate the implementation of our pipeline and suggest there might be at least some information within the estimated cortical microstructure that allows differentiation of subjects in terms of their prematurity status, although it is conceivable that the predictive results might be driven by some other confounding factors,

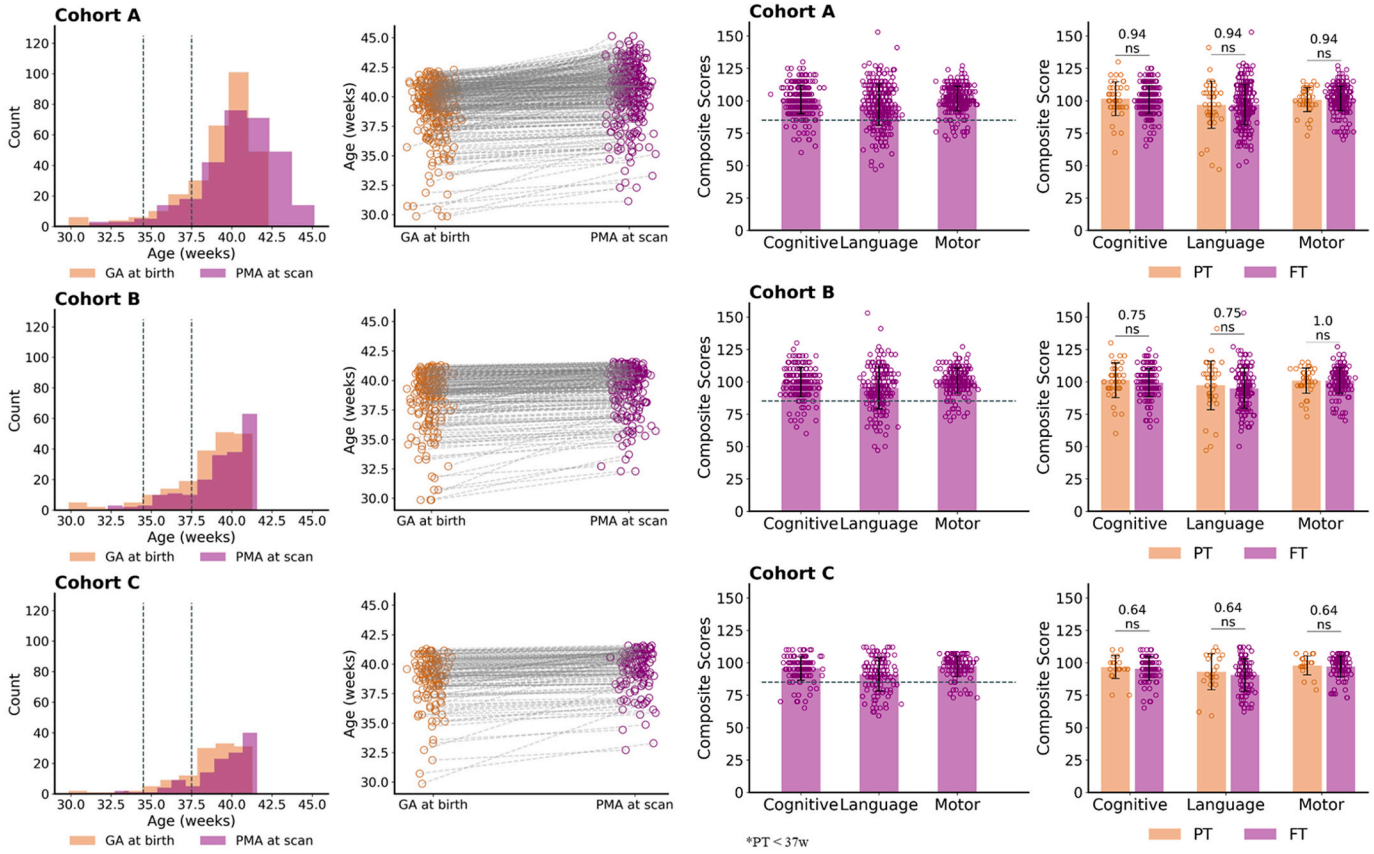


Fig. 4. Visual description of the Cohorts A,B and C showing distribution of GA at birth and PMA at scan as well as BSID-III composite scores.

such as motion (*Supp. Materials: Motion Correction section*). This information is however insufficient for the more difficult prediction of the continuous GA at birth.

3.3.2. *BSID-III score prediction*

Our results of the predictive models for BSID-III outcome suggested limited ability to distinguish Atypical vs Typical categories for cognitive, language, and motor composite scores (Fig. 5b). All models generally performed poorly, with the performance depending on the combination of the cohort and parcellation strategy. Once again, the random parcellation seemed to lead to better results supporting the idea there might be more granular information than that contained within 52 atlas parcels. However, the results were too heterogenous and close to random levels to make conclusions on optimal number of parcels with confidence. The erratic behaviour observed for Cohort D (especially for language and motor scores) might be due to the lower stability of the model to the small variations within the data caused by the randomly sampled small dataset, i.e. due to a less robust model which is unlikely to generalize well.

In contrast to the original study, the prediction results for the BSID-III continuous scores using cortical microstructure did not outcompete random levels (Supp. Materials: Predictive results, Fig. 5d). Given the results on the categorised scores, it is possible that some information pertaining to later developmental outcome is contained within the input data of early cortical microstructure, but was insufficient to predict the continuous scores like in the case of birth age prediction.

3.3.3. *Nested validation (hyperparameter tuning & independent test set)*

To expand on the validation strategy, we also performed a nested validation on all 4 predictive tasks: prematurity status, GA at birth, categorical BSID-III, and continuous BSID-III scores based on cortical features corrected for PMA at scan. This strategy enabled evaluation of mean metrics as well as calculation of confidence intervals for the predictive results (Table 3). Again, predicting prematurity status from the cortical microstructure was possible, but the more complex task of continuous GA at birth prediction was not successful. In the case of BSID-III scores, no model was better than random in neither categorised nor continuous setting.

To investigate whether an alternative model, incorporating information on the infant cortical microstructure with additional clinical risk factors and environmental information could lead to different results, we performed an additional prediction analysis using the global FA (corrected for PMA at scan), GA at birth, sex, and IMD score as inputs for the prediction of the prematurity status and categorised BSID-III scores. Despite the global effects (observed with ANCOVA) between the scores and the input variables within the Cohort A, such information did not lead to successful predictions within the nested validation setting (Supp. Table 5).

Additionally, because of a potential relationship between the whole-brain cortical FA metrics and motion parameters (Christiaens et al., 2021) (Supp. Table 6), we analyzed whether this confounder might have an impact on predictive results after the correction for PMA at scan. This was not the case (see Supp. Materials: Predictive Results, Supp. Table 7), suggesting that correction for motion might not be necessary in the context of this study.

3.4. *Overlap of input metrics*

We were also interested in evaluating whether the observed predictive results could be explained by analysis of the overlap of the inputs (global median FA as well as the regional FA of the atlas parcellations) during the classification considering only Cohort A as an example (Fig. 6). In short, we first estimated the distribution of FA medians over subjects per category (for prematurity status: PT/FT; for scores: Typical/

Table 2
Composition of Cohorts A, B, and C in terms of categorised BSID-III composite scores. In the *PT vs FT* columns, the percentages refer to percent of given PT (or FT) group from the entire PT (or FT) population within the cohort. |Δ%| is a relative difference between PT and FT % prevalence within the Atypical group.

	Atypical (scores ≤85)			Typical	
Cohort A (N = 295)	Number	PT vs FT	|Δ %|	Number	PT vs FT
Cognitive	24	PT: 4 (9.3%) FT: 20 (7.9%)	1.4	271	PT: 39 (90.7%) FT: 232 (92.1%)
Language	57	PT: 7 (16.3%) FT: 50 (19.8%)	3.5	238	PT: 36 (83.7%) FT: 202 (80.2%)
Motor	21	PT: 5 (11.6%) FT: 16 (6.3%)	5.3	274	PT: 38 (88.4%) FT: 236 (93.7%)
Cohort B (N = 196)					
Cognitive	20	PT: 4 (10.3%) FT: 16 (10.2%)	0.1	176	PT: 35 (89.7%) FT: 141 (89.8%)
Language	42	PT: 6 (15.4%) FT: 36 (22.9%)	7.5	154	PT: 33 (84.6%) FT: 121 (77.1%)
Motor	17	PT: 5 (12.8%) FT: 12 (7.6%)	5.2	179	PT: 34 (87.2%) FT: 145 (92.4%)
Cohort C (N = 126)					
Cognitive	16	PT: 2 (9.5%) FT: 14 (13.3%)	3.8	110	PT: 19 (90.5%) FT: 91 (86.7%)
Language	36	PT: 4 (19.0%) FT: 32 (30.5%)	11.5	90	PT: 17 (81.0%) FT: 73 (69.5%)
Motor	14	PT: 3 (14.3%) FT: 11 (10.5%)	6.2	112	PT: 18 (85.7%) FT: 94 (89.5%)

Atypical) and then quantified the measure of similarity between the two categories as the intersection of the areas under the two curves (normalized between 0 and 1). Overall, the overlap in terms of global cortical FA, after correction for PMA at scan, was around 80% for preterm vs term-born infants, which was lower than in the case of BSID-III scores where the FA overlap was above 85% for all scores. This overlap difference between categories might potentially explain the higher prediction of prematurity status compared with the BSID-III scores, i.e. the heterogeneity between BSID-III categories might not be sufficient to allow reliable prediction.

4. Discussion

In this study, we were unable to conceptually replicate the findings of Ouyang et al. (2020) that cortical microstructure at birth as assessed by DTI-based FA can reliably predict later neurodevelopmental outcome as evaluated with BSID-III in infants born preterm and full-term. Nevertheless, the performance on the prematurity task allowed us to draw limited conclusions: i) a sufficiently large and heterogeneous dataset appeared to be an indispensable factor for successful training and reliable validation of ML algorithms; ii) age corrections of the cortical FA features appeared to be required, and iii) random parcellations with higher number of parcels seemed beneficial compared with atlas based parcellations. Differences between the original study and our observations are discussed below in the light of deviations between methodologies and possible limitations.

4.1. Methodological considerations

In this replication study, we had to deviate from the original methodology used for regional FA extraction, mainly due to our inability to exactly re-implement the processing steps and software settings of the original study. Additionally, the potential replication using a different, but equally robust methodology would serve to strengthen the claim that cortical FA at birth might serve as a useful biomarker of later neurodevelopmental outcomes. In short, the diffusion data was preprocessed with correction for motion artefact and slice-to-volume reconstruction using the SHARD approach with all data passing the stringent dHCP quality control assessment (Edwards et al., 2022). To parcellate the cortex into 52 parcels based on an atlas (Feng et al., 2019), we first categorised all subjects into three template groups based on PMA at scan (33w,36w,39w). To register individual subject FA maps to a DTI-based relevant age-group atlas template, we opted to apply FSL's FNIRT following the recommendations of the atlas' authors (brainmrimap.org, n.d) whereas the original study used the LDDMM method (Miller et al., 2002). We then inverted the FNIRT computed transformation warps to bring cortical parcellations to individual subject space.

Instead of cortical FA extraction through cortical skeletons, we next projected the FA map and cortical parcellation for each subject to the individual inner cortical surface using a previously validated cylindrical approach, guided by the local minimum of axial diffusivity (AD) in the cortical ribbon (Lebenberg et al., 2019). This ensured that FA measures and parcel labels were extracted from the same spatial location at each voxel and in a reliable way at the individual level. Moreover, the current extraction method used in this study is robust across the wide range of ages at scan, since it relies on the local AD minimum which is consistently observed in the cortex at all ages rather than relying on FA measures which become low around term equivalent age (Batalle et al., 2019). Despite this, some errors in the subject parcellations were observed particularly in areas where the opposing gyral banks were in close spatial proximity in volumetric space for example at the level of the central sulcus. This was mainly due to the propagation of registration problems, perhaps related to the difference between PMA at scan and template age (i.e. lower registration quality when there was a larger age gap between a subject and the template). This might have particularly affected the predictive results of Cohort A which included more infants with ages distant to the 39w group template than other cohorts. To minimize the effects of the potential remaining errors of cortical parcellation after iterative corrections, we finally extracted the median rather than mean regional FA, due to its better robustness to outlier values particularly in small cortical regions (e.g. angular gyrus, parahipocampal gyrus). Together, we feel that this would mean it is unlikely that these methodological differences in our optimized pipeline for cortical FA extraction compared with the original study would have impacted the observation of robust prediction results, although this cannot be excluded.

Observed differences between extracted regional FA metrics across subjects and regions could stem from differences in infants' developmental stage (related to GA at birth and PMA at scan) and in intrinsic differences in local regional microstructure observed also in mature adult brains (Fukutomi et al., 2018). As both the original and this study focused on infants scanned close to birth, taking into account the inter-individual variability in cortical microstructure related to PMA at scan was necessary to reduce the impact of this potential confounding factor. In line with the original study and for the sake of simplicity, we then adjusted the regional cortical FA using a biphasic piecewise linear regression with inflection point at 36w PMA, although a region-dependent age cut-off might have been more appropriate to consider the spatially and temporally heterogenous maturation of

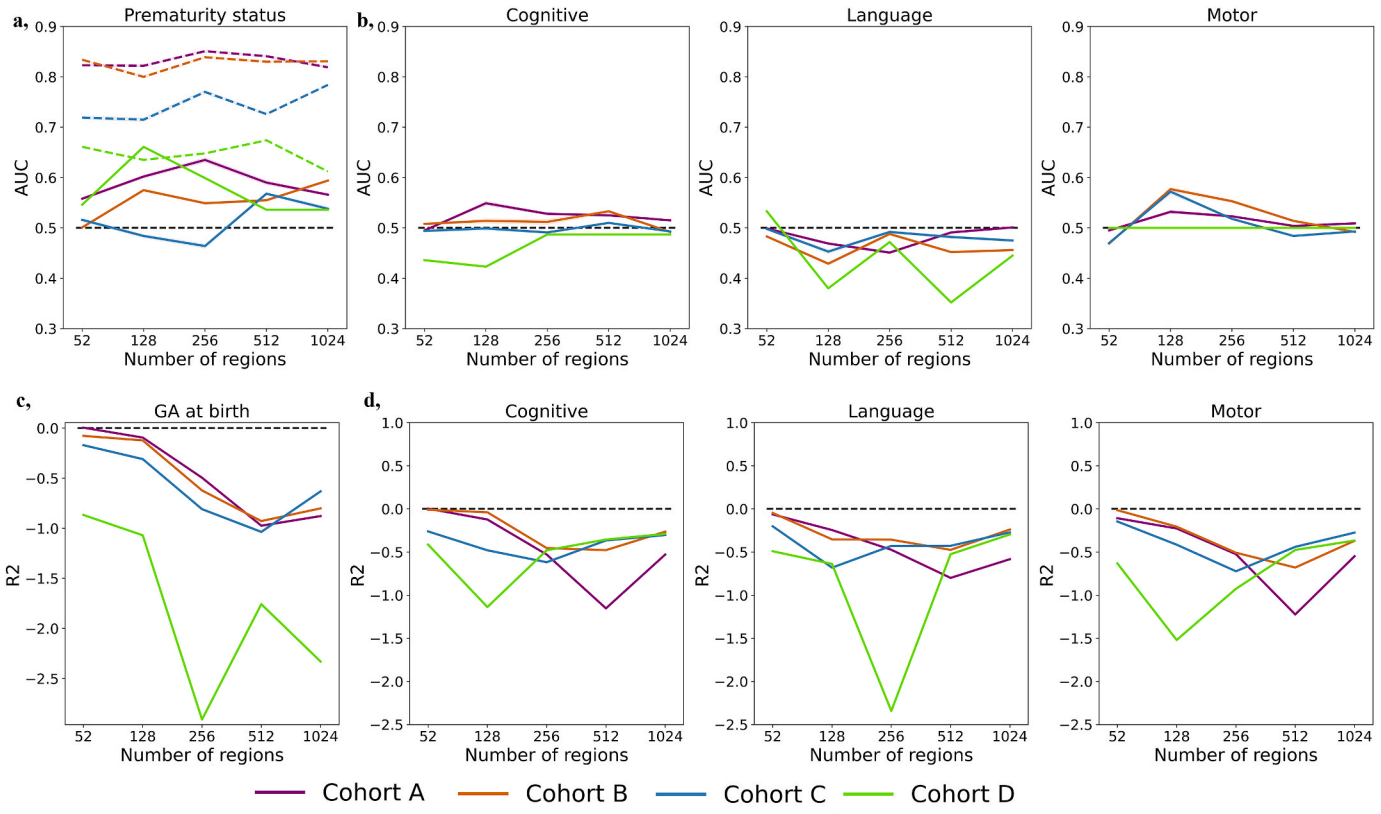


Fig. 5. Predictive results of categorical tasks (upper row) for **a,** prematurity status and **b,** categorised BSID-III scores with increasing region numbers across the Cohorts A, B, C and D, and the predictive results of the continuous tasks (lower row) for **c,** GA at birth and **d,** BSID-III scores. Coloured dotted lines in the prematurity status subplot show predictive results before the adjustment for PMA at scan. Grey dotter lines represent the random predictive level (0.5 for AUC and 0 for R2 metric).

cortical microstructure. This FA adjustment led to a substantial reduction in the prediction of prematurity status, suggesting that the initial results were driven mostly by differences in PMA at scan rather than in cortical microstructure per se. These results stressed the importance of appropriate corrections of inputs (such as PMA at scan) to dissociate the impact of cortical microstructure on predictive performance from the incidental effects of potential confounders. Reintroducing inter-regional differences (e.g. through regional value interpolation at TEA, Fig. 3c) could be done in the future to keep the inherent but potentially informative differences between the cortical parcels. For the same reason, it might also be worthwhile scaling all input FA features across regions simultaneously, as here (like in the original study), each input FA feature was scaled to lie in the range (0,1) over the group for each parcel independently.

Another important methodological aspect for prediction studies relies on the implemented pipeline. Results of ML studies are by nature, highly dependent on the dataset and its learning procedure, and on data splitting strategies. Approaches can be highly variable and thus results often generalize poorly to other datasets, especially in the context of small datasets (Woods, 2018; Varoquaux, 2018; Poldrack et al., 2020). In our case, even the largest Cohort A, which is six times larger than the dataset in the original study, would typically be considered small in a ML context and might also suffer from these generalization limitations. We attempted to remedy this by a thorough assessment of predictive performance which is a crucial step to dissociate whether the reported results are reliable or whether they might stem from performance fluctuations. Although the leave-one-out cross-validation strategy is popular for small datasets since it maximizes the training set size by keeping only a single subject for the testing, its main limitation is reduced estimation of generalization performance due to the depleted test set (Bouthillier et al., 2021). We thus aimed to leverage our larger cohort which allowed us to assess how, holding the algorithm constant, the performance changes under small perturbations to the data by performing randomized repetitions of the k-fold validation. Testing the prediction pipeline performance multiple times using multiple splits and thus test sets in this manner can improve estimation of performance variability (Poldrack et al., 2020; Bouthillier et al., 2021).

Finally, predictive performance is highly sensitive to hyperparameter settings with suboptimal settings leading to unfairly low performance (Poldrack et al., 2020). Noting that the hyperparameters used in the original study might be suboptimal in our input/target setting, we decided to incorporate the hyperparameter tuning within the nested cross-validation loop. However, neither the extended validation nor tuning led to behavioural outcome prediction results similar to those reported in the original study.

An important finding was that for prematurity status prediction, we observed significant differences in prediction performance in relation to the cortical parcellation used. One possible explanation is that summarising cortical microstructure into 52 regional metrics using a neonatal atlas might be insufficient for the behavioural predictions. We found that random parcellations led to better prediction of prematurity status, potentially simply due to the increased granularity of inputs associated with a larger number of parcels. However, there is also evidence that in the absence of a substantial sample size to train meaningful models, adding dimensions can negatively impact model performance and generalisability (Koutroumbas & Theodoridis, 2009). This is not only because of the 'curse of dimensionality', but also because more granular parcellations will be more sensitive to noise within the data and likely suffer from large multi-collinearity between the inputs (discussed below). Although outside the scope of our replication study, this suggested that future studies will clearly benefit from first identifying the study-specific optimal number of parcels that represent a good trade-off

Table 3
Results of the nested validation for **a,** categorical tasks and **b,** continuous tasks. CI: 95% confidence interval.

a, Categorical (corrected for PMA at scan)

Outcome	Inputs	AUC [CI]	SPEC [CI]	SENS [CI]
Prematurity status	ROIs (52)	0.571 [0.5279; 0.6148]	0.929 [0.8865; 0.9723]	0.213 [0.0862; 0.3404]
	Random (128)	0.564 [0.5135; 0.6145]	0.821 [0.7435; 0.8993]	0.307 [0.1569; 0.4564]
	Random (256)	0.562 [0.5025; 0.6223]	0.951 [0.9278; 0.9750]	0.173 [0.0599; 0.2868]
	Random (512)	0.502 [0.4341; 0.5694]	0.864 [0.8037; 0.9234]	0.140 [0.0185; 0.2615]
	Random (1024)	0.512 [0.4414; 0.5829]	0.864 [0.7723; 0.9564]	0.160 [0.0324; 0.2876]
Cognitive Score	ROIs (52)	0.474 [0.4340; 0.5138]	0.881 [0.8203; 0.9419]	0.067 [-0.0160; 0.1493]
	Random (128)	0.475 [0.4594; 0.4899]	0.916 [0.8858; 0.9462]	0.033 [-0.0080; 0.0747]
	Random (256)	0.478 [0.4657; 0.4900]	0.956 [0.9314; 0.9801]	0.000 [0.0000; 0.0000]
	Random (512)	0.498 [0.4944; 0.5006]	0.995 [0.9888; 1.0012]	0.000 [0.0000; 0.0000]
	Random (1024)	0.493 [0.4836; 0.5018]	0.985 [0.9672; 1.0035]	0.000 [0.0000; 0.0000]
Language Score	ROIs (52)	0.496 [0.4906; 0.5010]	0.945 [0.8778; 1.0131]	0.046 [-0.0111; 0.1034]
	Random (128)	0.460 [0.4261; 0.4930]	0.887 [0.7969; 0.9771]	0.032 [0.0077; 0.0564]
	Random (256)	0.479 [0.4525; 0.5051]	0.958 [0.9050; 1.0102]	0.000 [0.0000; 0.0000]
	Random (512)	0.491 [0.4793; 0.5025]	0.934 [0.8844; 0.9844]	0.047 [0.0089; 0.0860]
	Random (1024)	0.490 [0.4693; 0.5107]	0.920 [0.8725; 0.9684]	0.060 [0.0076; 0.1115]
Motor Score	ROIs (52)	0.512 [0.4789; 0.5446]	0.894 [0.8482; 0.9390]	0.130 [0.0270; 0.2330]
	Random (128)	0.501 [0.4713; 0.5302]	0.962 [0.9376; 0.9855]	0.040 [-0.0096; 0.0896]
	Random (256)	0.493 [0.4865; 0.4988]	0.985 [0.9731; 0.9976]	0.000 [0.0000; 0.0000]
	Random (512)	0.500 [0.5000; 0.5000]	1.000 [1.0000; 1.0000]	0.000 [0.0000; 0.0000]
	Random (1024)	0.498 [0.4948; 0.5006]	0.995 [0.9896; 1.0011]	0.000 [0.0000; 0.0000]

b, Continuous (corrected for PMA at scan)

Outcome	Inputs	RHO [CI]	MAE [CI]	R2 [CI]
GA at birth	ROIs (52)	0.092 [-0.0038; 0.1887]	1.464 [1.3303; 1.5976]	−0.069 [-0.1599; 0.0212]
	Random (128)	0.046 [-0.0915; 0.1834]	1.597 [1.4775; 1.7163]	−0.185 [-0.3618;-0.0073]
	Random (256)	−0.018 [-0.1568; 0.1206]	1.646 [1.4598; 1.8321]	−0.295 [-0.4497;-0.1411]
	Random (512)	−0.080 [-0.2404; 0.0809]	1.594 [1.4665; 1.7211]	−0.392 [-0.5363;-0.2480]
	Random (1024)	0.017 [-0.0970; 0.1316]	1.493 [1.3944; 1.5907]	−0.320 [-0.5269;-0.1121]
Cognitive Score	ROIs (52)	−0.026 [-0.0981; 0.0466]	6.613 [6.2772; 6.9489]	−0.081 [-0.1129;-0.0485]
	Random (128)	0.002 [-0.0476; 0.0508]	6.517 [6.2554; 6.7794]	−0.076 [-0.0959;-0.0567]
	Random (256)	−0.044 [-0.1063; 0.0177]	6.739 [6.5065; 6.9717]	−0.141 [-0.2019;-0.0804]
	Random (512)	0.043 [-0.0379; 0.1238]	6.461 [6.1927; 6.7291]	−0.045 [-0.0630;-0.0265]
	Random (1024)	0.065 [-0.0103; 0.1407]	6.430 [6.1602; 6.7006]	−0.043 [-0.0605;-0.0246]
Language Score	ROIs (52)	−0.072 [-0.1417;-0.0014]	10.500 [10.0920; 10.9080]	−0.079 [-0.1176;-0.0406]
	Random (128)	−0.060 [-0.1348; 0.0156]	10.535 [9.9543; 11.1153]	−0.079 [-0.1409;-0.0172]
	Random (256)	−0.011 [-0.0645; 0.0417]	10.822 [10.0066; 11.6369]	−0.137 [-0.2700;-0.0046]
	Random (512)	−0.099 [-0.1536;-0.0444]	10.778 [10.0128; 11.5437]	−0.154 [-0.2992;-0.0097]
	Random (1024)	−0.080 [-0.1296;-0.0309]	10.717 [10.0147; 11.4201]	−0.118 [-0.2220;-0.0149]
Motor Score	ROIs (52)	−0.206 [-0.2345;-0.1780]	6.074 [5.7814; 6.3665]	−0.074 [-0.1148;-0.0340]
	Random (128)	−0.045 [-0.0768;-0.0142]	6.013 [5.5622; 6.4639]	−0.065 [-0.0874;-0.0426]
	Random (256)	−0.019 [-0.0194;-0.0194]	6.061 [5.5185; 6.6033]	−0.083 [-0.1388;-0.0264]
	Random (512)	0 [0; 0]	5.865 [5.4868; 6.2437]	−0.031 [-0.0449;-0.0162]
	Random (1024)	0 [0; 0]	5.865 [5.4868; 6.2437]	−0.031 [-0.0449;-0.0162]

between information granularity and dimensionality reduction.

Another limitation to the atlas parcellations potentially stems from methodological difficulties regarding the delineation of cortical regions in the neonate cortex. This is particularly relevant for regions such as the hippocampus or entorhinal gyrus, which are small and difficult to delineate. As a result, FA values in these types of areas had substantially larger variability compared to other parcels. Such (heterogeneously) noisy measurements could thus negatively affect predictive performance. Although the benefits of atlas informed parcellations on predictive results should be investigated further in the future, our findings suggest random parcellations could provide a simpler and neutral strategy to extract cortical microstructure descriptors for use in predictive models. Alternatively, regularized voxel-based strategies combined with deep learning able to automatically extract the relevant

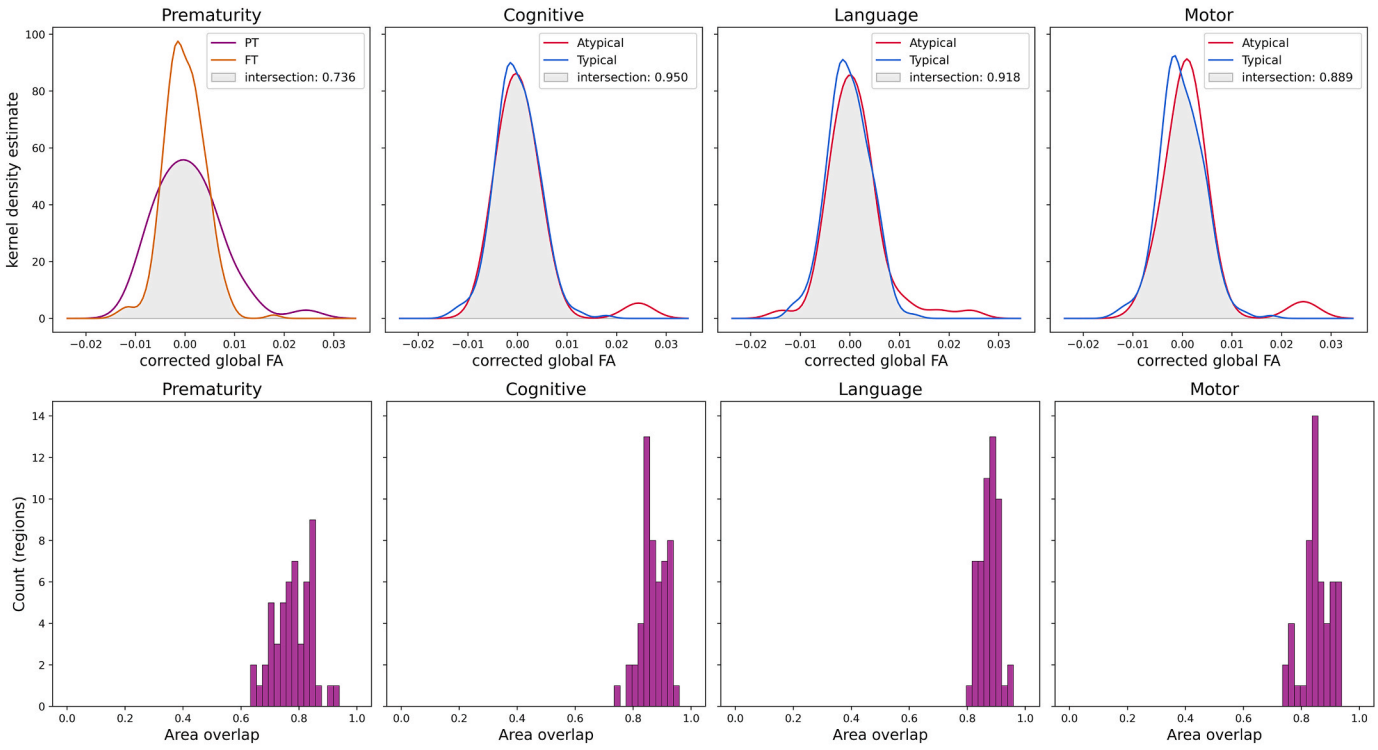


Fig. 6. Overlap of global FA PT and FT infants and between Atypical and Typical infants assessed by BSID-III scores (top row). Same assessment across 52 atlas parcels collapsed into a histogram (bottom row).

features from the large dataset images might be an additional avenue to explore to avoid the difficulties associated with the cortical parcellations.

4.2. *Possible causes for negative prediction results*

Our inability to recreate the conclusions of the original study might stem from deviations to the original methodology as discussed in the previous sections, as well as from differences in PMA at scan between cohorts and in image characteristics. For instance, the PMA distribution for Cohort A is skewed towards somewhat older ages, which is of relevance as the reliability of measures of cortical microstructure at these ages is still debated. Some authors have suggested that above 40w PMA, FA measures might be as low as noise values, and thus no longer sensitive to microstructural characteristics (Ball et al., 2013; Ouyang et al., 2019b), whereas others have proposed that FA provides relevant information even above 41w PMA (Batalle et al., 2019). Here, we observed inter-subject and inter-regional variability in the extracted FA metrics, even in the Cohort A where the skew towards the older subjects was the largest. This suggested that FA might hold some information about the underlying microstructural differences across subjects which, if useful for the establishment of relationship with the later neurodevelopment, should be able to drive the predictions. Nevertheless, given the ongoing discussion regarding FA's lack of sensitivity to microstructural changes in the cortex in older infants (>38w PMA), its usefulness for neurodevelopmental outcome prediction might be limited to the preterm cohorts.

Additionally, differences in the spatial resolution and signal-to-noise ratio of diffusion MRI images between studies might have led to different partial volume effects (critical at the level of the thin cortical ribbon of neonates) and variable estimation of DTI model and thus FA. However, the high quality of the dHCP dataset for anatomical and diffusion MRI has been highlighted by several studies from different groups. We are also confident about the reliability of our measures and analyses having observed successful prediction of prematurity status based on cortical FA (corrected for PMA at scan), even in reduced cohorts.

Besides, development of cortical microstructure is defined by multiple ongoing processes such as the growth of dendritic arborisation, synaptogenesis, and myelination of intracortical white matter, combined with the large-scale complex process of cortical folding. As in the original study, we focused on the widely used DTI-derived FA measure. In the preterm period, cortical FA is initially high due to the early radial orientation of apical dendrites and radial glia, and then decreases: which has been suggested to reflect increased neurite growth in all orientations as the cortex matures (McKinstry et al., 2002; Kroenke et al., 2007; Huang et al., 2008; Ball et al., 2013; Dubois et al., 2014; Smyser et al., 2016; Ouyang et al., 2019a; Ouyang et al., 2019b). However, the non-monotonic FA developmental profile in relation to PMA at scan suggests a more complex underlying biological picture where the later stages (around the age of full-term bir0h) may be dominated by increasing cellular and organelle density (Batalle et al., 2019). This complex relationship potentially makes FA a difficult input feature for linear predictive models requiring a robust age correction strategy. It is possible that a different DTI-derived metric with a simpler age-related profile, such as mean or axial diffusivity that continuously decreases with age, might be a more straightforward choice. Nevertheless, none of the DTI-derived metrics is likely to completely capture the diversity of properties of the underlying microstructure on its own. Thus, expansion of the input feature space to include the complementary DTI-derived metrics could be required. Such increase of the input space dimensions however would likely need to be combined with an appropriate feature selection strategy before (or during) the training of the predictive models (Gondová et al., 2022).

Moreover, the specificity of the DTI method might be limited even when combining metrics (Vos et al., 2012) and 'overlook' important cortical modifications that relate to functional outcomes. Using more elaborate diffusion MRI models, such as diffusion kurtosis imaging (DKI) (Jensen et al., 2005) or a multi-compartment neurite orientation dispersion and density imaging (NODDI) (Zhang et al., 2012) may improve description of complex cortical cytoarchitecture during

gestation (Genc et al., 2017; Mah et al., 2017; Kimpton et al., 2020). However, these complex models require multi-shell diffusion data which likely limits the future potential applications of predictive models with clinical applications. Although possible with the dHCP data used here, estimating these more complex models was beyond the scope this replication.

Additionally, cortical microstructure alone might provide too limited a view of the complex developmental stage that a brain has reached at birth. In this respect, the unimodal nature of diffusion imaging at birth might be unable to prognosticate about the complex combination of neurodevelopmental processes underlying later behavioural acquisitions. Assuming that information sufficient for the prediction of later outcomes relies mainly on the brain, incorporating other sources of information such as cortical morphology (Seidlitz et al., 2018; Fenchel et al., 2022), white matter connectivity (Ball et al., 2015; Wee et al., 2017; Girault et al., 2019), or even functional connectivity (He et al., 2021) might be beneficial. Nevertheless, timing of the imaging will have an important impact on what 'developmental information' can be accessed as the maturation occurs within a spatially heterogenous and temporally asynchronous progression (both within grey and white matter) (Kulikova et al., 2014; Croteau-Chonka et al., 2016; Gilmore et al., 2018; Lebenberg et al., 2019; Yu et al., 2019). It is also likely that developmental trajectories capturing the dynamism that characterises brain development at multiple timepoints can provide more reliable information for predictive models compared to a single snapshot of the brain at birth. Thus, investing into the acquisition of longitudinal cohorts, even if difficult to implement in practice, might be required for developing successful outcome prognostic tools in the future. As a more practical alternative at large scale, it might be useful to further investigate what is the optimal timing for a single MRI exploration to provide relevant markers for prediction.

A further consideration is that as children develop after birth, environmental information (e.g. sensory stimulation, nutrition, social interaction, socio-economic factors) also significantly impacts on neurodevelopment. This has been shown to explain a large part of observed interindividual variability in sensorimotor and cognitive skill acquisition, with the influence of perinatal risk factors diminishing over time (Pierrat et al., 2021; Boardman and Counsell, 2019; Gui et al., 2019; Beauregard et al., 2018; Mangin et al., 2017). Thus, it is unlikely that brain features at birth alone, in the absence of socio-economic and clinical information, can truly capture the complex relationships between a given child's development and outcomes. Adding these factors to predict outcome could thus provide useful complementary information. Nevertheless, given the high resource requirements for large population level MRI studies, it will be important to investigate whether the imaging data brings additional benefits to outcome prognostication compared to readily available clinical information.

Another aspect that might partly explain our negative prediction results is the low heterogeneity within our study population who were at relatively low risk of neurodevelopmental impairment. In contrast to the original study which used the BSID-III scores assessed around 2 years of corrected age, the neurodevelopmental outcome assessment available in the dHCP database was performed at around 18 months. The overall neurodevelopmental stage in terms of language, cognitive and motor acquisitions between the two ages is very similar and is unlikely to change the significance or conclusions derived from our results. Additionally, assessment at 18 months corresponds to a strategic age for identifying neurodevelopmental impairment in preterm infants, and thus examining whether cortical microstructure at birth relates to outcomes at this age is of interest even in low-risk preterm populations without obvious brain anomalies who can still develop subtle neuromotor disorders (e.g. developmental coordination disorder) (Groeschel et al., 2019; Spittle and Orton, 2014; Edwards et al., 2011; Zwicker et al., 2012). Although in our study, prematurely born infants were over-represented in the atypical category of the BSID-III motor score compared with full-terms, no major development delays or specific disabilities were neither expected nor observed within the specific low-risk cohort analyzed in our replication study. Therefore, an important question raised in the original study of insufficient heterogeneity within the input and outputs remained pertinent in our work.

As an extension, we performed the training and evaluation of the outcome predictions on different cohorts with expanded variability for the targets (Cohort B) as well as both targets and inputs (Cohort A), but without success in terms of prediction results. The limited variance within the input features combined with the relatively narrow distribution of the BSID-III scores across the dataset may have still been insufficient to allow the ML models to establish relationships with outcomes even if such relationships do exist. Moreover, cohort selection bias towards a low-risk population might remain problematic, given that even in the presence of inherent but subtle microstructural variability that could be predictive of the later outcomes, this might remain undetected because of the noisy metric extraction: detecting this relevant and predictive variability would thus require more precise measurements of cortical FA. Developing targeted models for a specific narrow age range may alleviate problems with feature extraction and provide more reliable inputs, which coupled with wider BSID-III outcome distributions (including more infants with atypical outcomes) could lead to more reliable results.

4.3. ML & interpretation

Finally, we would like to discuss limitations regarding the generalisability and interpretation of ML results, although we did not recreate the evaluation of model feature importance as in the original study, given the poor prediction of BSID-III scores we obtained. In the context of linear SVR, the high correlation between cortical features observed in our cohort (Supp. Fig. 5) and similar trends expected in the original study might be a critical issue to keep in mind. Firstly, input collinearity might affect the generalisability of the trained models. This is because during training, predictive models aim to minimize the prediction error, not to explain the true relationships between input features and the target. Thus, any of the highly correlated features can be selected as a strong predictor while the impact of other features with similar relationships with the target is minimized, leading to a lower performance on new data that does not follow the same correlation patterns.

In the future, it might be appropriate to attempt to decorrelate the inputs for all cortical parcels using a dimensionality reduction method such as PCA (Hotelling, 1933). Such a strategy might have additional benefits by decreasing the number of input dimensions and capturing only the components with the greatest amount of variance, while retaining most of the original information content and additionally "denoising" the data. Nevertheless, interpreting the feature weights in such a setting in terms of regional association with outcome suffers from the same pitfall. Post-hoc analysis of the feature weights will only allow evaluation of the use of features within the ML model, which is not the same as evaluating the data itself. Thus, our ability to draw links between the interpretation of predictive models and the underlying region-target functional associations will remain limited without additional hypothesis-driven experiments.

5. Conclusions

Machine learning has become a major focus of research as a promising tool for early behavioural outcome prediction. However, very few clinical questions, including the neurodevelopmental assessments, represent well posed discrimination tasks that can be naturally framed as ML problems (Baker and Kandasamy, 2022). Careful consideration of potential confounders that might drive predictions instead of the inputs of interest is required before ML can become useful in the clinical practice. This is further compounded by a lack of standardized methodologies for model implementation with some validation methods in current ML applications leading to highly variable and potentially

inflated ML performance estimation, in particular in small datasets (Varoquaux, 2018). As in many other areas, reproducibility failures in ML-based science appears prevalent without any systemic solutions (Kapoor and Narayanan, 2022). In light of these limitations, our inability to corroborate conclusions of the study by Ouyang et al. (2020) does not come as a surprise. Importantly, it should not be seen as a discouragement either. As a reminder, the normative charts used in clinical practice, even though proposing much simpler features for the prognosis, are often derived from very large cohorts. Successful leveraging of the complexity of brain imaging data now requires reaching (at minimum) similar large scales. Additionally, it might be useful to admit that replication problems, whether they involve direct or conceptual replication efforts, are an unescapable component of the current *early* stage of the field which could drive increased efforts to establish good practices for implementation and careful validation of the promising prediction tools.

Author contributions

Andrea Gondová: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **Sara Neumane:** Conceptualization, Investigation, Writing – review & editing**. Yann Leprince**: Conceptualization, Methodology, Software, Writing - review & editing. **Jean-François Mangin**: Conceptualization, Methodology, Software, Writing - review & editing. **Tomoki Arichi:** Conceptualization. Investigation, Resources, Writing - review & editing; **Jessica Dubois**: Conceptualization, Formal analysis, Investigation Methodology, Resources; Validation, Supervision, Writing - review & editing.

Funding

The developing Human Connectome Project was funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement no. 319456.

AG is supported by the CEA NUMERICS program. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 800945 — NUMERICS — H2020-MSCA-COFUND-2017.

SN was supported by a postdoctoral fellowship from the Bettencourt Schueller Foundation (www.fondationbs.org).

TA is supported by a MRC Clinician Scientist Fellowship [MR/P008712/1] and MRC Transition Support Award [MR/V036874/1].

JD received support from the Fondation Médisite (under the aegis of the Fondation de France, grant FdF-18-00092867) and the IdEx Université de Paris (ANR-18-IDEX-0001).

JFM and JD are supported by the Fondation Paralysie Cérébrale for the ENSEMBLE project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is freely available for download from www.developingconnectome.org

Acknowledgements

We thank Nicholas Harper for providing the individual clinical data; and the infants and their families for their participation in this study.

Abbreviations

AUC	Area Under the (Receiver Operating Characteristic) Curve
BSID-III	Bayley Scales of Infant and toddler Development, 3rd edition
dHCP	Developing Human Connectome Project
DTI	Diffusion tensor imaging
FA	fractional anisotropy
FT	full-term born infants
GA	gestational age
GM	grey matter
LOOCV	leave-one-out cross-validation
MAE	mean absolute error
ML	machine learning
MRI	magnetic resonance imaging
PMA	post-menstrual age
PC	principal component
PCA	Principal Component Analysis
PT	preterm born infants
R2	coefficient of determination
SENS	sensitivity
SPEC	specificity
SVR	support vector regression
WM	white matter

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ynirp.2023.100170.

References

April Adibpour, P., Lebenberg, J., Kabdebon, C., Dehaene-Lambertz, G., Dubois, J., 2020. Anatomo-functional correlates of auditory development in infancy. Devel. Cognit. Neurosci. 42, 100752.

Arpi, E., Ferrari, F., 2013. Preterm birth and behaviour problems in infants and preschool-age children: a review of the recent literature. Dev. Med. Child Neurol. 55.

Baker, S.B., Kandasamy, Y., 2022. Machine learning for understanding and predicting neurodevelopmental outcomes in premature infants: a systematic review. Pediatr. Res. 1–7.

Ball, G., Pazderová, L., Chew, A.T., Tusor, N., Merchant, N., Arichi, T., Counsell, S.J., 2015. Thalamocortical Connectivity Predicts Cognition in Children Born Preterm, vol. 25. Cerebral Cortex, New York, NY), pp. 4310–4318.

Ball, G., Srinivasan, L., Aljabar, P., Counsell, S.J., Durighel, G., Hajnal, J.V., Edwards, A. D., 2013. Development of cortical microstructure in the preterm human brain. Proc. Natl. Acad. Sci. USA 110, 9541–9546.

Batalle, D., O'Muircheartaigh, J., Makropoulos, A., Kelly, C.J., Dimitrova, R., Hughes, E. J., Counsell, S.J., 2019. Different patterns of cortical maturation before and after 38 weeks gestational age demonstrated by diffusion MRI in vivo. Neuroimage 185, 764–775.

Bayley, N., 2012. Bayley Scales of Infant and Toddler Development, third ed.

Beauregard, J.L., Drews-Botsch, C., Sales, J.M., Flanders, W.D., Kramer, M.R., 2018. Does socioeconomic status modify the association between preterm birth and children's early cognitive ability and kindergarten academic achievement in the United States? Am. J. Epidemiol. 187, 1704–1713.

Blauw-Hospers, C.H., de Graaf-Peters, V.B., Dirks, T., Bos, A.F., Hadders-Algra, M., 2007. Does early intervention in infants at high risk for a developmental motor disorder improve motor and cognitive development? Neurosci. Biobehav. Rev. 31, 1201–1212.

Bovzek, J., Makropoulos, A., Schuh, A., Fitzgibbon, S.P., Wright, R., Glasser, M.F., Robinson, E.C., 2018. Construction of a neonatal cortical surface atlas using multimodal surface matching in the developing human connectome project. Neuroimage 179, 11–29.

Boardman, J.P., Counsell, S.J., 2019. Factors Associated with Atypical Brain Development in Preterm Infants: Insights from Magnetic Resonance Imaging.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Vincent, P., 2021. Accounting for Variance in Machine Learning Benchmarks, 03098. *ArXiv, abs/2103*.

Christiaens, D., Cordero-Grande, L., Pietsch, M., Hutter, J., Price, A.N., Hughes, E.J., Tournier, J.D., 2021. Scattered slice SHARD reconstruction for motion correction in multi-shell diffusion MRI. Neuroimage 225.

Cordero-Grande, L., Hughes, E.J., Hutter, J., Price, A.N., Hajnal, J.V., 2018. Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: application to neonatal brain imaging. Magn. Reson. Med. 79, 1365–1376.

Croteau-Chonka, E.C., Dean, D.C., Remer, J., Dirks, H., O'Muircheartaigh, J., Deoni, S.C., 2016. Examining the relationships between cortical maturation and white matter myelination throughout early childhood. Neuroimage 125, 413–421.

Dubois, J., Dehaene-Lambertz, G., Kulikova, S., Poupon, C., Hüppi, P.S., Hertz-Pannier, L., 2014. The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants. Neuroscience 276, 48–71.
Edwards, A.D., Rueckert, D., Smith, S.M., Seada, S.A., Alansary, A., Almalbis, J.F., Hajnal, J.V., 2022. The developing human connectome project neonatal data release. Front. Neurosci. 16.
Edwards, J.K., Berube, M., Erlandson, K., Haug, S., Johnstone, H., Meagher, M., Zwicker, J.G., 2011. Developmental coordination disorder in school-aged children born very preterm and/or at very low birth weight: a systematic review. J. Dev. Behav. Pediatr. 32, 678–687.
Fenchel, D., Dimitrova, R., Robinson, E.C., Batalle, D., Chew, A.T., Falconer, S., O'Muircheartaigh, J., 2022. Neonatal multi-modal cortical profiles predict 18-month developmental outcomes. Devel. Cognit. Neurosci. 54.
Feng, L., Li, H., Oishi, K., Mishra, V.R., Song, L., Peng, Q., Huang, H., 2019. Age-specific gray and white matter DTI atlas for human brain at 33, 36 and 39 postmenstrual weeks. Neuroimage 185, 685–698.
Fukutomi, H., Glasser, M.F., Zhang, H., Autio, J.A., Coalson, T.S., Okada, T., Hayashi, T., 2018. Neurite imaging reveals microstructural variations in human cerebral cortical gray matter. Neuroimage 182, 488–499.
Genc, S., Malpas, C.B., Holland, S.K., Beare, R., Silk, T.J., 2017. Neurite density index is sensitive to age related differences in the developing brain. Neuroimage 148, 373–380.
Gilmore, J.H., Knickmeyer, R.C., Gao, W., 2018. Imaging structural and functional brain development in early childhood. Nat. Rev. Neurosci. 19, 123–137.
Girault, J.B., Munsell, B.C., Puechmaille, D., Goldman, B.D., Prieto, J.C., Styner, M., Gilmore, J.H., 2019. White matter connectomes at birth accurately predict cognitive abilities at age 2. Neuroimage 192, 145–155.
Gondová, A., Neumane, S., Leprince, Y., Mangin, J., Dubois, J., 2022. June 19). Infant cortical microstructure at term-equivalent age accurately predicts prematurity status at birth. In: OHBM 2022 Annual Meeting. Glasgow, UK.
Groeschel, S., Holmström, L., Northam, G.B., Tournier, J.D., Baldeweg, T., Latal, B., Vollmer, B., 2019. Motor abilities in adolescents born preterm are associated with microstructure of the corpus callosum. Front. Neurol. 10.
Gui, L., Loukas, S., Lazeyras, F., Huppi, P.S., Meskaldji, D.E., Borradori-Tolsa, C., 2019. Longitudinal study of neonatal brain tissue volumes in preterm infants and their ability to predict neurodevelopmental outcome. Neuroimage 185, 728–741.
Hadaya, L., Nosarti, C., 2020. The neurobiological correlates of cognitive outcomes in adolescence and adulthood following very preterm birth. Semin. Fetal Neonatal Med., 101117
He, L., Li, H., Chen, M., Wang, J., Altaye, M., Dillman, J.R., Parikh, N.A., 2021. Deep multimodal learning from MRI and clinical data for early prediction of neurodevelopmental deficits in very preterm infants. Front. Neurosci. 15.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 498–520.
Huang, H., Yamamoto, A., Hossain, M.A., Younes, L., Mori, S., 2008. Quantitative cortical mapping of fractional anisotropy in developing rat brains. J. Neurosci. 28, 1427–1433.
Hughes, E.J., Winchman, T., Padormo, F., Teixeira, R.P., Wurie, J., Sharma, M., Hajnal, J.V., 2017. A dedicated neonatal brain imaging system. Magn. Reson. Med. 78, 794–804.
Hutter, J., Tournier, J.D., Price, A.N., Cordero-Grande, L., Hughes, E.J., Malik, S.J., Hajnal, J.V., 2018. Time-efficient and flexible design of optimized multishell HARDI diffusion. Magn. Reson. Med. 79, 1276–1292.
Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K., 2005. Diffusional kurtosis imaging: the quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. Magn. Reson. Med. 53.
Johnston, M.V., 2009. Plasticity in the developing brain: implications for rehabilitation. Develop. Disabil. Res. Rev. 15 2, 94–101.
Kapoor, S., Narayanan, A., 2022. Leakage and the reproducibility crisis in ML-based science. In: Leakage and the Reproducibility Crisis in ML-based Science (arXiv).
Kimpton, J.A., Batalle, D., Barnett, M.L., Hughes, E.J., Chew, A.T., Falconer, S., Counsell, S.J., 2020. Diffusion magnetic resonance imaging assessment of regional white matter maturation in preterm neonates. Neuroradiology 63, 573–583.
Kostović, I., Sedmak, G., Judas, M., 2019. Neural histology and neurogenesis of the human fetal and infant brain. Neuroimage 188, 743–773.
Kroenke, C.D., Essen, D.C., Inder, T.E., Rees, S., Bretthorst, G.L., Neil, J.J., 2007. Microstructural changes of the baboon cerebral cortex during gestational development reflected in magnetic resonance imaging diffusion anisotropy. J. Neurosci. 27, 12506–12515.
Kulikova, S., Hertz-Pannier, L., Dehaene-Lambertz, G., Buzmakov, A., Poupon, C., Dubois, J., 2014. Multi-parametric evaluation of the white matter maturation. Brain Struct. Funct. 220, 3657–3672.
Lebenberg, J., Mangin, J.F., Thirion, B., Poupon, C., Hertz-Pannier, L., Leroy, F., Dubois, J., 2019. Mapping the asynchrony of cortical maturation in the infant brain: a MRI multi-parametric clustering approach. Neuroimage 185, 641–653.
Mah, A., Geeraert, B.L., Lebel, C.A., 2017. Detailing neuroanatomical development in late childhood and early adolescence using NODDI. PLoS One 12.
Makropoulos, A., Gousias, I.S., Ledig, C., Aljabar, P., Serag, A.M., Hajnal, J.V., Rueckert, D., 2014. Automatic whole brain MRI segmentation of the developing neonatal brain. IEEE Trans. Med. Imag. 33, 1818–1831.
Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S.P., Bo\vzek, J., Rueckert, D., 2018. The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112.
Mangin, K.S., Horwood, L.J., Woodward, L.J., 2017. Cognitive development trajectories of very preterm and typically developing children. Child Dev. 88 1, 282–298.
Marín, O., 2016. Developmental timing and critical windows for the treatment of psychiatric disorders. Nat. Med. 22, 1229–1238.
McKinstry, R.C., Mathur, A.M., Miller, J.H., Ozcan, A., Snyder, A.Z., Schefft, G.L., Neil, J. J., 2002. Radial organization of developing preterm human cerebral cortex revealed by non-invasive water diffusion anisotropy MRI. Cerebr. Cortex 12 12, 1237–1243.
Miller, M.I., Trouvé, A., Younes, L., 2002. On the metrics and euler-Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. 4, 375–405.
Müller, A.B., Saccani, R., Valentini, N.C., 2017. Impact of compensatory intervention in 6- to 18-month-old babies at risk of motor development delays. Early Child. Dev. Care 187, 1707–1717.
Oishi, K., Mori, S., Donohue, P.K., Ernst, T., Anderson, L., Buchthal, S., Chang, L., 2011. Multi-contrast human neonatal brain atlas: application to normal neonate development analysis. Neuroimage 56, 8–20.
Ouyang, M., Dubois, J., Yu, Q., Mukherjee, P., Huang, H., 2019a. Delineation of early brain development from fetuses to infants with diffusion MRI and beyond. Neuroimage 185, 836–850.
Ouyang, M., Jeon, T., Sotiras, A., Peng, Q., Mishra, V.R., Halovanic, C., Huang, H., 2019b. Differential cortical microstructural maturation in the preterm human brain with diffusion kurtosis and tensor imaging. Proc. Natl. Acad. Sci. U. S. A 116, 4681–4688.
Ouyang, M., Peng, Q., Jeon, T., Heyne, R., Chalak, L., Huang, H., 2020. Diffusion-MRI-based regional cortical microstructure at birth for predicting neurodevelopmental outcomes of 2-year-olds. Elife 9.
Parikh, N.A., 2016. Advanced neuroimaging and its role in predicting neurodevelopmental outcomes in very preterm infants. Semin. Perinatol. 40 8, 530–541.
Pickler, R.H., Mcgrath, J.M., Reyna, B.A., McCain, N.L., Lewis, M., Cone, S., Best, A.M., 2010. A model of neurodevelopmental risk and protection for preterm infants. J. Perinat. Neonatal Nurs. 24, 356–365.
Pierrat, V., Marchand-martin, L., Marret, S., Arnaud, C., Benhammou, V., Cambonie, G., Ancel, P.-Y., 2021. Neurodevelopmental outcomes at age 5 among children born preterm: EPIPAGE-2 cohort study. BMJ 373.
Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of Best Practices for Evidence for Prediction: A Review. JAMA Psychiatry 77 (5), 534–540.
April Rolland, C., Lebenberg, J., Leroy, F., Moulton, E., Adibpour, P., Riviere, D., Dubois, J., 2019. Exploring microstructure asymmetries in the infant brain cortex: a methodological framework combining structural and diffusion mri. In: 2019 IEEE 16th International *Symposium On Biomedical Imaging*. ISBI 2019, pp. 426–429.
Seidlitz, J., Vá\vsa, F., Shinn, M., Romero-García, R., Whitaker, K.J., Vértes, P.E., Bullmore, E.T., 2018. Morphometric similarity networks detect microscale cortical organization and predict inter-individual cognitive variation. Neuron 97, 231–247. e7.
Silbereis, J.C., Pochareddy, S., Zhu, Y., Li, M., Sestan, N., 2016. The cellular and molecular landscapes of the developing human central nervous system. Neuron 89, 248–268.
Smyser, C.D., Dosenbach, N.U., Smyser, T.A., Snyder, A.Z., Rogers, C.E., Inder, T.E., Neil, J.J., 2016. Prediction of brain maturity in infants using machine-learning algorithms. Neuroimage 136, 1–9.
Smyser, T.A., Smyser, C.D., Rogers, C.E., Gillespie, S.K., Inder, T.E., Neil, J.J., 2016. Cortical gray and adjacent white matter demonstrate synchronous maturation in very preterm infants. Cerebr. Cortex 26 8, 3370–3378.
Spittle, A.J., Orton, J., 2014. Cerebral palsy and developmental coordination disorder in children born preterm. Semin. Fetal Neonatal Med. 19 2, 84–89.
Theodoridis, S., Koutroumbas, K.D., 2008. Pattern Recognition, fourth ed.
Tournier, J.D., Christiaens, D., Hutter, J., Price, A.N., Cordero-Grande, L., Hughes, E.J., Hajnal, J.V., 2019. A data-driven approach to optimising the encoding for multi-shell diffusion MRI with application to neonatal imaging. NMR Biomed. 33 e4348 - e4348.
Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77.
Vos, S.B., Jones, D.K., Jeurissen, B., Viergever, M.A., Leemans, A., 2012. The influence of complex white matter architecture on the mean diffusivity in diffusion tensor MRI of the human brain. Neuroimage 59, 2208–2216.
Wee, C.-Y., Tuan, T.A., Broekman, B.F., Ong, M.Y., Chong, Y.-S., Kwek, K., Qiu, A., 2017. Neonatal neural networks predict children behavioral profiles later in life. Hum. Brain Mapp. 38.
Woods, B., 2018. Expanding Search in the Space of Empirical ML, 01495. *ArXiv, abs/1812*.
Yu, Q., Ouyang, A., Chalak, L., Jeon, T., Chia, J.M., Mishra, V.R., Huang, H., 2016. Structural development of human fetal and preterm brain cortical plate based on population-averaged templates. Cerebr. Cortex 26 11, 4381–4391.
Yu, Q., Peng, Y., Kang, H., Peng, Q., Ouyang, M., Slinger, M., Huang, H., 2019. Differential White Matter Maturation from Birth to 8 Years of Age, vol. 30. Cerebral Cortex, New York, NY), pp. 2674–2690.
Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C., 2012. NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 61, 1000–1016.
October 20 Zwaan, R., Etz, A., Lucas, R., Donnellan, B., 2017. Making Replication Mainstream, vol. 41. Behavioral and Brain Sciences.
Zwicker, J.G., Yoon, S.W., Mackay, M., Petrie-Thomas, J., Rogers, M., Synnes, A.R., 2012. Perinatal and neonatal predictors of developmental coordination disorder in very low birthweight children. Arch. Dis. Child. 98, 118–122.