Language Function Following Preterm Birth: prediction using Machine Learning

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

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

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

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

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

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

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

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

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

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

***Corresponding author:** Correspondence concerning this article should be addressed to Evdoxia Valavani, Usher Institute, Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. Email: evdoxia.valavani@ed.ac.uk. Tel: +30 6983756291.

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

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

Disclosure statement: The authors have no conflicts of interest relevant to this article to disclose.

Category of study: Clinical study

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

Impact:

- A combination of clinical perinatal factors and neonatal DTI measures of white matter microstructure leads to accurate prediction of language outcome at 2 years corrected gestational age following preterm birth.
- A model that comprises clinical and MRI features that has potential to be scalable across centres. It offers a basis for enhancing the power and generalizability of

diagnostic and prognostic studies of neurodevelopmental disorders associated with language impairment.

- Early identification of infants who are at risk of language delay, facilitating targeted early interventions and support services, which could improve the quality of life for children born preterm.

Abstract

Background

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

Methods

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

Results

The model achieved balanced accuracy:91%, sensitivity:86%, and specificity:96%. The probability of language delay at two years CGA is increased with: increasing values of peak width of skeletonized fractional anisotropy (PSFA), radial diffusivity (PSRD), and axial diffusivity (PSAD) derived from dMRI; among twins; and after an incomplete course of, or no exposure to, antenatal corticosteroids. Female sex and breastfeeding during the neonatal period reduced the risk of language delay.

Conclusion

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

Introduction

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

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

Nonetheless, a few studies have built and validated tools for prediction of the composite outcome of neurodevelopmental impairment at 2 years corrected gestational age (CGA) for children born preterm. Tyson et al. 2008[10] investigated the clinical and demographic characteristics of a cohort of infants born before 26 weeks of gestation and found that the risk of adverse neurodevelopmental outcome at 18 to 22 months CGA was predicted using gestational age (GA), sex, exposure to antenatal corticosteroids, multiple birth and birth weight.

Ambalavanan et al. 2012[11] reported that neurodevelopmental impairment at 18 to 22 months CGA was predicted by combining sex, respiratory illness severity, and enlarged ventricular size, periventricular leukomalacia or porencephalic cyst on cranial ultrasound. Vesoulis et al. 2018[12] developed a tool for prediction of risk of neurodevelopmental impairment at 18 to 24 months CGA. This tool comprised ventilator days, mode of delivery, exposure to antenatal corticosteroids, retinopathy of prematurity (ROP) requiring surgery, and magnetic resonance imaging (MRI) findings (cerebellar haemorrhage size, cerebellar haemorrhage laterality, intraventricular haemorrhage grade, white matter injury).

However, deficits in different developmental domains require different therapies and targeted support strategies. Thus, tools for stratification of children at high risk of impairment in specific developmental domains would be valuable. Recently, Vassar et al. 2020[13] evaluated the predictive value of structural MRI and DTI variables for classification of very preterm infants at high versus low risk of language delay. They developed a model for prediction of language delay that included DTI variables in three brain regions and achieved 89% sensitivity and 86% specificity. Ball et al. 2017[14] revealed that distinct patterns of brain structure and microstructure following preterm birth are linked to specific clinical and environmental factors, and these patterns correlate with neurodevelopmental outcome at 18 to 24 months CGA. Language outcome was associated with specific neuroanatomic variation, which was linked to: age at scan, need for continuous positive airway pressure, birth weight, GA at birth, parenteral nutrition, surfactant administration, and mechanical ventilation.

In view of this evidence, we hypothesized that a combination of clinical, environmental and imaging factors derived from DTI that capture generalised white matter dysmaturation would potentially enhance the prediction of language outcomes at 2 years CGA following preterm birth. Blesa et al. 2020[15] demonstrated that histogram-based variables derived from DTI (peak

width of skeletonized [PS] -FA, -MD, -RD, and -AD), which represent generalised water content and myelination, can be used as biomarkers of microstructural white matter alterations associated with preterm birth. The advantage of the histogram-based framework is that it is fully automated, captures generalized white matter dysmaturation which characterizes the encephalopathy of prematurity, is computationally inexpensive compared with tract-specific approaches, and has high inter-scanner reproducibility[16].

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

Methods

Participants

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

Clinical and Demographic Features

The selection of clinical and demographic features included in models was guided by extant literature linking biological and environmental exposures with neurocognitive development in preterm infants. Specifically, we studied the contribution towards prediction of language outcome at two years CGA of the following features: sex[10,11,18,19], GA (based on first trimester ultrasound)[10,18], birth weight[10,20], maternal age[21], primiparity[19], twin status[10,20], maternal body mass index (BMI)[22], medical history of maternal depression[23], administration of a complete course of antenatal corticosteroids for fetal lung maturation (defined as two doses 24 hours apart), any antenatal corticosteroid exposure[10,12,19,20], administration of antenatal magnesium sulphate ($MgSO_4$) for neuroprotection[24], mode of delivery (spontaneous vaginal delivery [SVD] or caesarean section)[19], total days requiring intubation whilst in the neonatal intensive

care unit (NICU)[11,12,18], bronchopulmonary dysplasia (BPD, defined as oxygen requirement at ≥36 weeks CGA)[19,20,25,26], late onset sepsis (LOS, defined as blood stream infection occurring ≥72 hours postnatally with (a) bacterial pathogen isolated from blood culture, or (b) blood culture growing coagulase negative staphylococcus, along with one or more signs of generalized infection, and treatment with intravenous antibiotics for 5 or more days)[20], necrotizing enterocolitis (NEC, defined as stages two or three according to the modified Bell's staging for NEC[27])[25,28], ROP treated with laser therapy[12,29], and type of infant feeding at discharge from the neonatal unit (dichotomized as exclusive maternal breast milk versus exclusive formula or mixed feeding)[30]. All infants had placental histopathology performed and histological chorioamnionitis was defined using an established system[31]. Maternal level of education (dichotomized as secondary school or below versus college, university or postgraduate studies)[18–20], and socioeconomic status of the family, operationalised as Scottish Index of Multiple Deprivation 2016 (SIMD16) quintile, where 1 indicates the most deprived and 5 indicates the least deprived (https://www2.gov.scot/Topics/Statistics/SIMD), were also included.

Image Acquisition

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

A Siemens MAGNETOM Verio 3-Tesla MRI clinical scanner (Siemens Healthcare Gmbh, Erlangen, Germany) and 12-channel phased-array head coil were used to acquire dMRI data consisting of 11 T2-and 64 diffusion-weighted (b=750 s/mm^2) single-shot, spin-echo, echo

planar imaging volumes collected in the axial plane with 2 mm isotropic voxels (TR=7300 ms, TE= 06 ms, FOV=256 mm, acquired matrix =128×128, 50 contiguous interleaved slices with 2 mm thickness, acquisition time=9 min 29 s).

Image Analysis

For each participant the dMRI was denoised using a Marchenko-Pastur-PCA-based algorithm[32,33]; eddy current and head movement were corrected using outlier replacement[34–36] and bias field inhomogeneity correction was performed by calculating the bias field of the mean b0 volume and applying the correction to all the volumes[37]. For each participant, PSFA, PSMD, PSRD, PSAD were calculated using age-optimized methods described by Blesa et al. 2020[15]. In summary, image data were registered to the Edinburgh Neonatal $Atlas_{50}$ (ENA_{50})[15] using a tensor registration[38], and their DTI maps were calculated. Subsequently, the individual FA maps were projected into the template skeleton and multiplied by the atlas custom mask. Finally, the peak width of the histogram values within the skeletonized maps was calculated as the difference between the 95^{th} and 5^{th} percentiles[39]. Figure 1 illustrates a summary of the process described. The code necessary to calculate histogram based metrics can be found at https://git.ecdf.ed.ac.uk/jbrl/psmd. Figure 2 shows scatterplots of the values of the peak width of skeletonized DTI metrics for all participants.

Language Outcome

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

outcome; children whose score was below 85 were considered to have moderate to severe language impairment, while scores equal to or greater than 85 were considered as normal-range or higher[41].

Data Analysis

We compared three feature selection algorithms: (a) Boruta[42], (b) ReliefF expRank[43,44], and (c) Random forests (RF) variable importance[45]. The Boruta algorithm is a wrapper feature selection technique built around the random forests learner, which uses Z score as the importance measure. In other words, it measures the importance of each feature by dividing the average loss of accuracy among all trees by the standard deviation of the accuracy loss. The basic idea of the ReliefF algorithm is to assign a 'weight' value to all features of a dataset based on how well their values distinguish between the instances that are near to each other and thus, how useful they are in predicting the response variable. The important features will have a large weight, while the redundant ones will have a low weight. In random forests variable importance, variable importance is computed using the mean decrease in Gini index. We can measure the total amount that the Gini index is decreased by splits over a given feature, averaged over all trees. A large value indicates an important feature. In all cases, we obtain a feature ranking indicating in descending order their contribution towards prediction of the response variable. The final feature subset for each feature selection algorithm was selected using leave-one-out cross-validation (LOOCV), using only the training dataset in each cross-validation iteration and following the process described by Tsanas et al. 2012[46]. Subsequently, the selected feature subset was presented into a RF classifier[47] in order to predict the binarized language composite score. Partial dependence plots (PDP)[48] were constructed in order to assess how the selected features influence the prediction of the RF classifier. To quantify the strength of the association between the

selected features, we used correlation analysis (the Spearman's rank correlation coefficient was used to quantify the strength of the association between two continuous features, the phi coefficient was used to quantify the association between two binary features, and the point-biserial correlation coefficient was used to quantify the strength of the association between a continuous and a binary feature).

The dataset is imbalanced since only 16% of the study group had a language composite score below 85. To overcome the class imbalance problem in the dataset, we explored different data balancing techniques; under-sampling of the majority class, over-sampling of the minority class, and the synthetic minority over-sampling technique (SMOTE)[49], which has been previously used in similar unbalanced applications in the healthcare domain[50–54]. We found that SMOTE yields the best results, which are presented in the paper. SMOTE is a training data enrichment method, where the minority class is over-sampled by creating new synthetic samples, to create a balanced dataset. For each minority class sample, the k minority class nearest neighbours were identified (using the suggestion of Chawla et al. with k=5) and synthetic samples were introduced along the line segments joining any or all of the k minority class nearest neighbours. Model validation was implemented using LOOCV. LOOCV involves holding out a single observation to be used as the test set, while the learner is trained using the remaining *n*-1 observations (*n* is the total number of observations). The process is repeated *n* times and each time a different observation from the original dataset is used as the test set. The result is *n* estimates of the test error. The final test error rate is the average of these *n* test error estimates. The accuracy of the model was assessed by constructing a confusion matrix which is a contingency table of the observed and predicted classes. Missing data for both numeric and categorical features were imputed using multiple imputation by chained equations (five imputed datasets were created in each LOOCV iteration)[55,56], based only on the information in the training set independently within each

LOOCV iteration. Data analysis was conducted in R. The R packages used were: tidyverse, dplyr, caret, randomForest, CORElearn, Boruta, mice, ggplot2, DMwR, Hmisc, RGraphics, grid, gridExtra, gridGraphics.

Results

Two-year language data and dMRI of the brain at term equivalent age were available from 89 children; demographic and clinical characteristics of the study population are presented in Table 1. At median age 24.13 months (range 23.24-28.27 months), 14 children had a language composite score below 85. The percentage of missing values in the dataset was 0.2% (one participant had missing histological chorioamnionitis data, two participants had missing SIMD16 and three participants had missing maternal BMI).

Figure 3 illustrates the out-of-sample performance of the RF classifier (trained on approximately 150 samples in each LOOCV iteration) as a function of the number of features selected by the different feature selection algorithms. These data show that feeding a subset of eight features selected by the Boruta feature selection algorithm (a wrapper feature selection technique built around the RF learner) to the RF classifier gives the highest balanced accuracy. The selected feature subset comprises PSFA, twin status (yes or no), antenatal steroid exposure (complete or incomplete course), any antenatal steroid exposure (yes or no), sex (male or female), PSRD, PSAD, and feeding at discharge from the NICU (exclusive maternal breast milk versus exclusive formula or mixed feeding). Figure 4 shows the importance attributed to each feature by each of the feature selection algorithms. PSFA, twin status, the course of antenatal steroid exposure, any antenatal steroid exposure, sex, PSRD, PSAD, and feeding are the jointly most predictive features towards the prediction of the binarized language outcome. PDP were used to visualize relationships between the selected features and the response based on our model (see Figure 5). The PDP provide insight into the effect of changing one or two features in terms of the model's prediction (binary response variable, indicating whether language composite score <85). Regarding the histogram-based variables derived from DTI, the PDP show that the predicted language impairment probability rises with increasing PSFA, PSRD, and PSAD values. PSRD and PSAD are presented in the same plot because they are

highly correlated as illustrated in the correlogram and correlation matrix in Figure 6. Language composite score <85 at 2 years CGA is more likely following a twin pregnancy, an incomplete course of antenatal corticosteroids, or no exposure to antenatal steroids. Female sex and feeding with exclusive breast milk reduce the risk of future language delay.

Table 2 shows the confusion matrix of the out-of-sample classification performance of the RF classifier when mapping the selected feature subset (i.e., PSFA, twin status, antenatal corticosteroid exposure, sex, PSRD, PSAD, and feeding at discharge) to the binarized language composite score. Our model achieved balanced accuracy: 91%, sensitivity: 86%, and specificity: 96%.

Finally, we repeated the analysis to investigate separately the performance of the model when presented only with either clinical or MRI features, which led to reduced model performance. As shown in Table 3, the model that comprises clinical and MRI features outperformed the models using only clinical or MRI features. The combination of clinical and DTI features enhances the prediction of language outcomes at 2 years CGA following preterm birth.

Discussion

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

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

importance is an embedded method. The use of ReliefF and the RF importance have been extensively used and validated in many different applications and we have previously conducted a thorough empirical study[57] where they performed very competitively against many established feature selection approaches. In general, we would expect a wrapper or embedded method to perform better for a particular choice of a classifier, although it might not necessarily generalize very well with the choice of different classifiers."

Our findings suggest that PSFA, PSRD, and PSAD, which detect generalized white matter microstructural alterations in preterm infants compared to infants born at term (15), are predictive of impaired language development at two years CGA. We explored the predictive value of whole brain measures of peak width of skeletonized DTI metrics, instead of tract specific segmentations, because preterm brain dysmaturation is a substantially generalized process[58], and language development draws on broad cognitive capacities. We have found that the probability of language delay is higher with increased PSFA, PSRD, and PSAD. These features are consistent with delayed myelination, less coherent white matter organization, and altered axonal integrity in the preterm brain[15,59]. Previous research has also shown that abnormalities in brain structure following preterm birth are correlated with long-term neurodevelopmental outcome[60].

The data show that twin status is associated with increased risk of impaired language development. This finding is consistent with studies in the extant literature which have found that multiple pregnancy is associated with neurodevelopmental impairment[10,20,61] and language delay[62] at 2 years CGA. Language delay in twins can be attributed to postnatal environmental factors[63,64]; twins receive a less focused and less elaborated communicative interchange with their parents than do singletons. Thorpe et al. 2003[63] compared families with twins to families

with pairs of closely spaced singletons. This study found that language delay in twins compared to singletons may be explained by patterns of parent-child interaction and communication. Antenatal corticosteroid administration is associated with lower risk of language deficits, which has been previously proved by research[10,12]. Our findings suggest that male sex is a risk factor for language impairment in early childhood, consistent with previous studies which have associated male sex with poorer neurodevelopmental outcome following preterm birth[10,11,18,19]. Moreover, previous work has shown that exclusive breast milk feeding in the weeks following preterm birth can enhance brain development[30], and in the general population breast milk intake in infancy is associated with improved performance on intelligence tests[65]. In line with this, we found that exclusive breastfeeding is associated with improved language outcomes compared to formula feeding or mixed breast and formula feeding. It is surprising that GA at birth was not included in the final feature set. However, its influence on long-term outcome may be captured by PSRD and PSAD which are strongly correlated with GA at birth[15].

This study is the first to investigate the use of peak width of skeletonized DTI metrics as predictors for language development in the preterm population. The advantage of using these image biomarkers is that their calculation is fully automated, computationally inexpensive, and has high inter-scanner reproducibility[39], meaning that they can be easily obtained for preterm neonates who undergo a dMRI scan at term-equivalent age, and can be used for multi-centre studies. Thus, our model comprises features which can be easily obtained for future clinical application.

Hitherto, few studies have focused on developing and validating prediction models for early neurodevelopmental outcomes for children born preterm. Most tools predict the composite outcome of neurodevelopmental impairment[10–12]. However, deficits in different developmental domains require different interventions. Therefore, tools for timely identification of children at

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

The main strength of our study is that we had a longitudinal cohort of preterm infants that is deeply phenotyped with brain imaging and biological information that enabled us to investigate a large number of clinical, demographic, social, and DTI variables. We acknowledge some limitations in our study. The sample size is relatively small, and this is a single centre study so despite our best efforts with standard model validation techniques to assess model generalization we would need to further validate findings in a different cohort. Nonetheless, the study population was fairly representative of NICU populations in terms of comorbidities that have been associated with long-term neurodevelopmental outcomes. In addition, cortical grey matter was not assessed in this study. We focused on alterations in white matter microstructure, since it is the most consistently abnormal finding in preterm infants, by measuring a functionally tractable property using a tool that is readily applied to clinical image data. Future studies could aim to validate our model in additional external cohorts, and also apply machine learning techniques for prediction of motor, cognitive and social-emotional outcomes for children born preterm.

Conclusion

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

References

1. Chawanpaiboon, S. *et al.* Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob. Heal.* **7**, e37–e46 (2019).
2. Pierrat, V. *et al.* Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in France in 2011: EPIPAGE-2 cohort study. *BMJ* **358**, j3448 (2017).
3. van Noort-van der Spek, I. L., Franken, M.-C. J. P. & Weisglas-Kuperus, N. Language functions in preterm-born children: a systematic review and meta-analysis. *Pediatrics* **129**, 745–54 (2012).
4. Law, J., Rush, R., Schoon, I. & Parsons, S. Modeling Developmental Language Difficulties From School Entry Into Adulthood: Literacy, Mental Health, and Employment Outcomes. *J. Speech, Lang. Hear. Res.* **52**, 1401–1416 (2009).
5. Conti-Ramsden, G., Mok, P. L. H., Pickles, A. & Durkin, K. Adolescents with a history of specific language impairment (SLI): Strengths and difficulties in social, emotional and behavioral functioning. *Res. Dev. Disabil.* **34**, 4161–4169 (2013).
6. Spittle, A., Orton, J., Anderson, P. J., Boyd, R. & Doyle, L. W. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database Syst. Rev.* CD005495 (2015). doi:10.1002/14651858.CD005495.pub4
7. Linsell, L., Malouf, R., Morris, J., Kurinczuk, J. J. & Marlow, N. Prognostic Factors for Poor Cognitive Development in Children Born Very Preterm or With Very Low Birth Weight: A Systematic Review. *JAMA Pediatr.* **169**, 1162–72 (2015).
8. Linsell, L., Malouf, R., Morris, J., Kurinczuk, J. J. & Marlow, N. Prognostic factors

for cerebral palsy and motor impairment in children born very preterm or very low birthweight: a systematic review. *Dev. Med. Child Neurol.* **58**, 554–69 (2016).

9. Feldman, H. M., Lee, E. S., Yeatman, J. D. & Yeom, K. W. Language and reading skills in school-aged children and adolescents born preterm are associated with white matter properties on diffusion tensor imaging. *Neuropsychologia* **50**, 3348–62 (2012).
10. Tyson, J. E. *et al.* Intensive care for extreme prematurity--moving beyond gestational age. *N. Engl. J. Med.* **358**, 1672–81 (2008).
11. Ambalavanan, N. *et al.* Outcome trajectories in extremely preterm infants. *Pediatrics* **130**, e115-25 (2012).
12. Vesoulis, Z. A., El Ters, N. M., Herco, M., Whitehead, H. V & Mathur, A. M. A Web-Based Calculator for the Prediction of Severe Neurodevelopmental Impairment in Preterm Infants Using Clinical and Imaging Characteristics. *Child. (Basel, Switzerland)* **5**, (2018).
13. Vassar, R. *et al.* Neonatal Brain Microstructure and Machine-Learning-Based Prediction of Early Language Development in Children Born Very Preterm. *Pediatr. Neurol.* (2020). doi:10.1016/J.PEDIATRNEUROL.2020.02.007
14. Ball, G. *et al.* Multimodal image analysis of clinical influences on preterm brain development. *Ann. Neurol.* **82**, 233–246 (2017).
15. Blesa, M. *et al.* Peak Width of Skeletonized Water Diffusion MRI in the Neonatal Brain. *Front. Neurol.* **11**, 235 (2020).
16. Baykara, E. *et al.* A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. *Ann. Neurol.* **80**, 581–92 (2016).
17. Boardman, J. P. *et al.* Impact of preterm birth on brain development and long-term outcome: protocol for a cohort study in Scotland. *BMJ Open* **10**, e035854 (2020).

18. Charkaluk, M. L. *et al.* Neurodevelopment of children born very preterm and free of severe disabilities: the Nord-Pas de Calais Epipage cohort study. *Acta Paediatr.* **99**, 684–9 (2010).

19. Wood, N. S. *et al.* The EPICure study: associations and antecedents of neurological and developmental disability at 30 months of age following extremely preterm birth. *Arch. Dis. Child. Fetal Neonatal Ed.* **90**, F134-40 (2005).

20. Vohr, B. R., Wright, L. L., Poole, W. K. & McDonald, S. A. Neurodevelopmental outcomes of extremely low birth weight infants <32 weeks' gestation between 1993 and 1998. *Pediatrics* **116**, 635–43 (2005).

21. Tseng, K.-T. *et al.* The impact of advanced maternal age on the outcomes of very low birth weight preterm infants. *Medicine (Baltimore).* **98**, e14336 (2019).

22. Reynolds, L. C., Inder, T. E., Neil, J. J., Pineda, R. G. & Rogers, C. E. Maternal obesity and increased risk for autism and developmental delay among very preterm infants. *J. Perinatol.* **34**, 688–92 (2014).

23. Bozkurt, O. *et al.* Does maternal psychological distress affect neurodevelopmental outcomes of preterm infants at a gestational age of ≤32weeks. *Early Hum. Dev.* **104**, 27–31 (2017).

24. Marret, S. *et al.* [Effect of magnesium sulphate on mortality and neurologic morbidity of the very-preterm newborn (of less than 33 weeks) with two-year neurological outcome: results of the prospective PREMAG trial]. *Gynecol. Obstet. Fertil.* **36**, 278–88 (2008).

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

26. Twilhaar, E. S. *et al.* Cognitive Outcomes of Children Born Extremely or Very Preterm Since the 1990s and Associated Risk Factors: A Meta-analysis and Meta-

regression. *JAMA Pediatr.* **172**, 361–367 (2018).

27. Bell, M. J. *et al.* Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. *Ann. Surg.* **187**, 1–7 (1978).
28. van Vliet, E. O. G., de Kieviet, J. F., Oosterlaan, J. & van Elburg, R. M. Perinatal infections and neurodevelopmental outcome in very preterm and very low-birth-weight infants: a meta-analysis. *JAMA Pediatr.* **167**, 662–8 (2013).
29. Schmidt, B., Davis, P. G., Asztalos, E. V., Solimano, A. & Roberts, R. S. Association Between Severe Retinopathy of Prematurity and Nonvisual Disabilities at Age 5 Years. *JAMA* **311**, 523 (2014).
30. Blesa, M. *et al.* Early breast milk exposure modifies brain connectivity in preterm infants. *Neuroimage* **184**, 431–439 (2019).
31. Anblagan, D. *et al.* Association between preterm brain injury and exposure to chorioamnionitis during fetal life. *Sci. Rep.* **6**, 37932 (2016).
32. Veraart, J., Fieremans, E. & Novikov, D. S. Diffusion MRI noise mapping using random matrix theory. *Magn. Reson. Med.* **76**, 1582–1593 (2016).
33. Tournier, J.-D. *et al.* MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* **202**, 116137 (2019).
34. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208–S219 (2004).
35. Andersson, J. L. R. & Sotiropoulos, S. N. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* **125**, 1063–1078 (2016).
36. Andersson, J. L. R., Graham, M. S., Zsoldos, E. & Sotiropoulos, S. N. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. *Neuroimage* **141**, 556–572 (2016).

37. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–20 (2010).

38. Zhang, H., Yushkevich, P. A., Alexander, D. C. & Gee, J. C. Deformable registration of diffusion tensor MR images with explicit orientation optimization. *Med. Image Anal.* **10**, 764–785 (2006).

39. Baykara, E. *et al.* A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. *Ann. Neurol.* **80**, 581–92 (2016).

40. Albers, C. A. & Grieve, A. J. Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development– Third Edition. San Antonio, TX: Harcourt Assessment. *J. Psychoeduc. Assess.* **25**, 180–190 (2007).

41. Johnson, S., Moore, T. & Marlow, N. Using the Bayley-III to assess neurodevelopmental delay: which cut-off should be used? *Pediatr. Res.* **75**, 670–674 (2014).

42. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the **Boruta** Package. *J. Stat. Softw.* **36**, 1–13 (2010).

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

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

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

46. Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J. & Ramig, L. O. Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. *IEEE Trans. Biomed. Eng.* **59**, 1264–1271 (2012).

47. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

48. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).

49. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

50. Dessie, E. Y., Tsai, J. J. P., Chang, J.-G. & Ng, K.-L. A novel miRNA-based classification model of risks and stages for clear cell renal cell carcinoma patients. *BMC Bioinformatics* **22**, 270 (2021).

51. Park, K. H., Batbaatar, E., Piao, Y., Theera-Umpon, N. & Ryu, K. H. Deep learning feature extraction approach for hematopoietic cancer subtype classification. *Int. J. Environ. Res. Public Health* **18**, 1–24 (2021).

52. Lee, Y. W., Choi, J. W. & Shin, E. H. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med.* **129**, (2021).

53. Ivanović, M. D. *et al.* Predicting defibrillation success in out-of-hospital cardiac arrested patients: Moving beyond feature design. *Artif. Intell. Med.* **110**, (2020).

54. Nguyen, Q. D. N., Liu, A. B. & Lin, C. W. Development of a neurodegenerative disease gait classification algorithm using multiscale sample entropy and machine learning classifiers. *Entropy* **22**, 1–1818 (2020).

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

56. Van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16**, 219–242 (2007).

57. Tsanas, A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. (2012).

58. Telford, E. J. *et al.* A latent measure explains substantial variance in white matter microstructure across the newborn human brain. *Brain Struct. Funct.* **222**, 4023–4033 (2017).

59. Boardman, J. P. & Counsell, S. J. Invited Review: Factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. *Neuropathol. Appl. Neurobiol.* **46**, 413–421 (2020).

60. Batalle, D., Edwards, A. D. & O'Muircheartaigh, J. Annual Research Review: Not just a small adult brain: understanding later neurodevelopment through imaging the neonatal brain. *Journal of Child Psychology and Psychiatry and Allied Disciplines* **59**, 350–371 (2018).

61. Wadhawan, R. *et al.* Twin gestation and neurodevelopmental outcome in extremely low birth weight infants. *Pediatrics* **123**, e220-7 (2009).

62. Adams-Chapman, I., Bann, C. M., Vaucher, Y. E., Stoll, B. J. & Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Association between feeding difficulties and language delay in preterm infants using Bayley Scales of Infant Development-Third Edition. *J. Pediatr.* **163**, 680-5.e1–3 (2013).

63. Thorpe, K., Rutter, M. & Greenwood, R. Twins as a natural experiment to study the causes of mild language delay: II: Family interaction risk factors. *J. Child Psychol. Psychiatry* **44**, 342–355 (2003).

64. Thorpe, K. Twin children's language development. *Early Human Development* **82**, 387–395 (2006).

65. Horta, B. L., Loret De Mola, C. & Victora, C. G. Breastfeeding and intelligence: A systematic review and meta-analysis. *Acta Paediatrica, International Journal of Paediatrics* **104**, 14–19 (2015).

Figure 1. Scheme of the steps necessary for the calculation of the peak width of skeletonized DTI metrics. First, participants are registered to a template, then skeletonized and multiplied by a mask to calculate the histogram.

Figure 2. Scatterplots of the PSFA, PSMD, PSAD, and PSRD values for all participants. PSFA: peak width of skeletonized fractional anisotropy; PSMD: peak width of skeletonized mean diffusivity; PSAD: peak width of skeletonized axial diffusivity; PSRD: peak width of skeletonized radial diffusivity.

Figure 3. Comparison of out-of-sample LOOCV balanced accuracy results of the random forests classifier using the features selected by each of the three feature selection algorithms.

Figure 4. Feature importance plots. A) Importance attributed to each feature by the Boruta algorithm. The first eight features coloured in blue (PSFA, twin status, course of antenatal steroids, any antenatal steroids, sex, PSRD, PSAD, feeding at discharge) are the jointly most predictive features towards the prediction of language outcome. B) Importance attributed to features by RF variable importance. C) Importance attributed to features by ReliefF expRank. Computation of feature importance depends on the feature selection algorithm used, and is expressed relative to the maximum.

Figure 5. Partial dependence plots for the eight features selected by Boruta and used in the random forests classifier. A) The predicted language impairment probability rises with increasing PSFA values. B) 3D plot of PSRD and PSAD. The predicted language impairment

probability rises with increasing PSRD, and PSAD values. C) A twin pregnancy increases the predicted probability of language impairment. D) An incomplete course of antenatal corticosteroids increases the predicted probability of language impairment. E) No exposure to any antenatal steroids increases the predicted probability of language impairment. F) Female sex reduces the predicted probability of language impairment. G) Feeding with exclusive breast milk reduce the predicted probability of language impairment.

Language composite score <85 at 2 years CGA is more likely following a twin pregnancy, an incomplete course of antenatal corticosteroids, or no exposure to any antenatal steroids. Female sex and feeding with exclusive breast milk reduce the risk of future language delay.

Figure 6. Correlogram and correlation matrix of the eight most important features selected by the Boruta algorithm; p<.05 '*', p<.01 '**', p<.001 '***', p<.0001 '****'.