Predicting 2-year neurodevelopmental outcomes in extremely preterm infants using graphical network and machine learning approaches

Sandra E. Juul,[a,d] Thomas R. Wood,[a,d,*] Kendell German,[a] Janessa B. Law,[a] Sarah E. Kolnik,[a] Mihai Puia-Dumitrescu,[a] Ulrike Mietzsch,[a] Semsa Gogcu,[b] Bryan A. Comstock,[c] Sijia Li,[c] Dennis E. Mayock,[a] and Patrick J. Heagerty,[c] on behalf of the PENUT Consortium

[a]Division of Neonatology, Department of Pediatrics, University of Washington, Seattle, WA, USA
[b]Division of Neonatology, Department of Pediatrics, Wake Forest School of Medicine, NC, USA
[c]Department of Biostatistics, University of Washington, Seattle, WA, USA

eClinicalMedicine
2023;56: 101782

Published Online XXX
https://doi.org/10.1016/j.eclinm.2022.101782

Summary

Background Infants born extremely preterm (<28 weeks' gestation) are at high risk of neurodevelopmental impairment (NDI) with 50% of survivors showing moderate or severe NDI when at 2 years of age. We sought to develop novel models by which to predict neurodevelopmental outcomes, hypothesizing that combining baseline characteristics at birth with medical care and environmental exposures would produce the most accurate model.

Methods Using a prospective database of 692 infants from the Preterm Epo Neuroprotection (PENUT) Trial, which was carried out between December 2013 and September 2016, we developed three predictive algorithms of increasing complexity using a Bayesian Additive Regression Trees (BART) machine learning approach to predict both NDI and continuous Bayley Scales of Infant and Toddler Development 3rd ed subscales at 2 year follow-up using: 1) the 5 variables used in the National Institute of Child Health and Human Development (NICHD) Extremely Preterm Birth Outcomes Tool, 2) 21 variables associated with outcomes in extremely preterm (EP) infants, and 3) a hypothesis-free approach using 133 potential variables available for infants in the PENUT database.

Findings The NICHD 5-variable model predicted 3–4% of the variance in the Bayley subscale scores, and predicted NDI with an area under the receiver operator curve (AUROC, 95% CI) of 0.62 (0.56–0.69). Accuracy increased to 12–20% of variance explained and an AUROC of 0.77 (0.72–0.83) when using the 21 pre-selected clinical variables. Hypothesis-free variable selection using BART resulted in models that explained 20–31% of Bayley subscale scores and AUROC of 0.87 (0.83–0.91) for severe NDI, with good calibration across the range of outcome predictions. However, even with the most accurate models, the average prediction error for the Bayley subscale predictions was around 14–15 points, leading to wide prediction intervals. Higher total transfusion volume was the most important predictor of severe NDI and lower Bayley scores across all subscales.

Interpretation While the machine learning BART approach meaningfully improved predictive accuracy above a widely used prediction tool (NICHD) as well as a model utilizing NDI-associated clinical characteristics, the average error remained approximately 1 standard deviation on either side of the true value. Although dichotomous NDI prediction using BART was more accurate than has been previously reported, and certain clinical variables such as transfusion exposure were meaningfully predictive of outcomes, our results emphasize the fact that the field is still not able to accurately predict the results of complex long-term assessments such as Bayley subscales in infants born EP even when using rich datasets and advanced analytic methods. This highlights the ongoing need for long-term follow-up of all EP infants.

Funding Supported by the National Institute of Neurological Disorders and Stroke U01NS077953 and U01NS077955.

Keywords: Extreme prematurity; prediction; Outcomes; neurodevelopmental impairment

*Corresponding author. Division of Neonatology, Department of Pediatrics, University of Washington Medical Center, Box 356320, Seattle, WA, 98195, USA.
E-mail address: tommyrw@uw.edu (T.R. Wood).

[d]Contributed equally.

Research in context

Evidence before this study

Neurodevelopmental impairment (NDI) remains more prevalent in extremely preterm (EP) survivors than their term-born peers, with an increased risk of developmental delay, cerebral palsy, deafness, blindness as well as behavioral and psychological disorders, all of which impact independent functioning. Predicting outcomes of EP infants has become increasingly important for both clinicians and parents as survival improves. Recent summaries of the literature have identified dozens of prediction algorithms; however, these generally ignore socioeconomic factors and do not account for non-linear relationships or interactions between variables. For example, the most widely used prediction tool, the National Institute of Child Health and Human Development (NICHD) Extremely Preterm Birth Outcomes Tool, uses only five simple variables available at birth (sex, gestational age, birthweight, singleton pregnancy, and antenatal steroid exposure) and focuses largely on risk of mortality and a dichotomous NDI outcome.

Added value of this study

We attempted to overcome limitations of previous outcome prediction models using multiple methods. We predicted outcomes as both dichotomous severe NDI as well as continuous outcome in complex outcome assessments. Our machine learning approach, BART, allows for non-linear associations between predictors and outcome, and our network analyses consider how all the variables are associated with one-another before examining which variables are significantly associated with the outcome. We used 10-fold cross-validation for all predictions, and all models were assessed with multiple methods, including assessment of discrimination and calibration. Prediction variables also included basic information on socioeconomic factors and the home environment. As a result, predictions were meaningfully improved relative to prior strategies although continuous outcome prediction retained a relatively large range of individual prediction uncertainty.

Implications of all the available evidence

By using advanced analytic methods and a range of clinical and demographic predictor variables, most of which could be easily abstracted from medical records, outcome prediction in EP survivors becomes more accurate. These approaches could be used to significantly improve upon current outcome prediction tools. Though meaningful dichotomous NDI prediction may become possible, predicting complex continuous neurodevelopmental outcomes remains a challenge. Several in-hospital factors that clinicians may be able to use to improve in-hospital care and NDI risk prediction were identified, but a dominant effect of the home and wider environment on long-term outcome is also becoming increasingly clear and is supported by our data.

Introduction

Over 10% of babies born in the U.S. are preterm (<37 weeks' gestation), and over 95% of those survive.[1,2] Short-term outcomes of babies born extremely preterm (EP, <28 weeks' gestation) have improved over the past decades, with more infants surviving to hospital discharge without experiencing intraventricular hemorrhage (IVH), necrotizing enterocolitis (NEC), bronchopulmonary dysplasia (BPD), severe retinopathy of prematurity (ROP), or sepsis.[3–5] However, this decrease in short-term morbidities has not translated to improvements in neurodevelopmental outcomes for EP infants. In a recent study of 10,877 infants born at 22–28 weeks of gestation, of the 2458 surviving infants born at 22–26 weeks of gestation who were followed-up at 2 years corrected age (CA), 49% had no or mild neurodevelopmental impairment (NDI), 29% had moderate, and 21% had severe NDI.[6] A recent study of temporal trends also suggested that the number of EP infants surviving with major disabilities has remained relatively stable for the last 30 years.[7]

The presence of one or more morbidities (IVH, NEC, BPD, ROP and sepsis) has been associated with NDI, as has low gestational age (GA), male sex, and low birth weight or being small for gestational age (SGA).[8,9] Using data collected for the Preterm Erythropoietin (Epo) Neuroprotection Trial (PENUT), we recently published additional clinical factors associated with lower Bayley Scales of Infant and Toddler Development 3rd ed. (BSID-III) scores at 22–30 months CA: more packed red blood cell (pRBC) transfusions,[10] lower iron supplementation,[11] longer duration of sedation with opioids or benzodiazepines,[12] and longer duration of dexamethasone treatment.[13] How these complications of prematurity and treatment decisions interact to affect 2-year outcomes is not well understood. Dozens of outcome prediction algorithms have also been identified in the literature; however, these generally ignore socioeconomic factors, do not account for non-linear relationships or interactions between variables, and generally develop and validate their predictive models in the same infants.[14]

The PENUT Trial enrolled 936 EP infants born at 24-0/7 to 27-6/7 weeks' gestation with follow up assessments done at 2-years CA. Using the PENUT database, we now develop and cross-validate predictive models try to improve our ability to predict neurodevelopmental outcomes by applying two novel statistical modelling approaches - graphical network analysis and machine learning using Bayesian Additive Regression Trees (BART). These approaches consider whether individual variables interact with one another and identify which factors have the most effect on cognitive, motor, and language BSID-III scores. We were particularly interested in modifiable factors that might change physician/caretaker behavior as well as demographic and other factors that might identify infants at greater risk of NDI.

Methods

Data source and study population

All infants enrolled in the PENUT Trial (NCT #01378273) who survived and were assessed for long-term developmental outcome were eligible for this study.[15] The PENUT Trial was approved by an institutional review board (IRB) at each site. Parental consent was obtained prior to infant enrollment. We collected data about maternal characteristics, pregnancy, and delivery, as well as infant characteristics including exposure to medications and comorbidities during their NICU stay. At 20–33 months CA, infants were evaluated by certified examiners who assessed cognitive, motor, and language development with the BSID-lll. All BSID-III subscales were scaled based on the age which the assessment was performed, as is standard for the BSID-III. The population of interest consisted of infants who received at least one BSID-lll subscale assessment. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines. Completed STROBE and TRIPOD checklists are included in the supplementary material.

Statistical analysis

Demographic data

We used descriptive statistics to describe the demographic and baseline maternal and infant characteristics of the n = 692 included infants who had complete data for at least one BSID-III subscale, separated by those who did or did not have severe NDI (Table 1). Of these infants, n = 625 (90.3%) were assessed at prespecified PENUT assessment timepoint of 22–26 months CA,[15] with a further n = 67 also included who were assessed at 20–33 months CA. In PENUT, severe NDI was defined as the presence of severe cerebral palsy or a BSID-III composite motor score or composite cognitive score of less than 70.[15]

Graphical network analysis

To examine the relationships between variables known to be associated with long-term outcome, as well as their net effect on each of the of the BSID-III subscales, we used graphical network analysis. Graphical network analysis involves extracting significant relationships from a precision matrix of inter-related variables, which allows for the identification of important relationships after taking into account how all the other variables are related. When temporal relationships are known, this method can also allow for the identification of likely causal pathways that may include multiple nodes. To determine significant relationships in the network we used the method described by Williams and Rast.[16] The precision matrix was constructed using a maximum likelihood estimation (MLE) method and significant relationships were determined using Fisher Z-transformed 95% confidence intervals, which results in stable networks in scenarios where the number of predictors is fewer than the number of observations.[16]

The MLE methodology was applied using 21 predictors previously identified as being potentially independently associated with outcomes in EP infants. These variables were based on the five variables used in the NICHD EP Outcomes Prediction Tool – GA in weeks, birthweight, sex, prenatal steroids, and single vs multiple gestation pregnancy. However, we specified SGA rather than birthweight and at least 2+ doses of steroids, as these appear to be independently associated with outcomes. We also included sick appearance at birth, lowest hematocrit during admission, cumulative pRBC transfusion volume, total oral iron dose from birth to 36 weeks postmenstrual age (PMA), 5 min Apgar score <5, spontaneous intestinal perforation (SIP), pulmonary hemorrhage, culture proven symptomatic sepsis, Bell's stage 2b or 3 NEC,[17] grade III or IV IVH,[18] severe ROP (stage ≥3), severe BPD (supplemental oxygen requirement at 36 weeks PMA), prolonged exposure (>7 days) to opioids and/or benzodiazepines,[12] prolonged exposure (>14 days) to dexamethasone,[13] and maternal education (as a three-level variable of high school or less, some college, or Bachelor's degree or higher). Epo treatment in PENUT was also included as it is known to interact with iron status and transfusion requirements.[10,11] Of the 21 included variables, three variables had at least one infant missing data (n = 2–4 per variable). These were all binary variables (SGA, sick appearance at birth, 5 min Apgar <5), which were imputed by assuming that the factor was not present (e.g., "no" to yes/no questions). Significant associations in the resulting precision matrices were depicted as interconnected network diagrams using the ggraph, igraph, and qgraph libraries in R (Version 4.1.2, Foundation for Statistical Computing, Vienna, Austria).[19]

To complement the graphical network analysis, for each BSID-III subscale we constructed a linear generalized estimating equation (GEE) regression model using robust standard errors and an independence covariance structure using the same 21 predictors as used in the networks.[20] The GEE model structure was used to account for potential correlation of outcomes for same-birth siblings. From each BSID-III subscale

	Normal-to-moderate NDI	Severe NDI
Total included Infants, n (% of total)	614	78
Maternal and demographic characteristics		
Gestational age at birth, mean (SD)	25.6 (1.1)	25.1 (1.1)
Birth weight (g), mean (SD)	824 (187)	746 (175)
Small for gestational age, n (%)	74 (12.1)	19 (24.4)
Maternal education, n (%)		
High school or less	193 (31.4)	28 (35.9)
Some college	192 (31.3)	20 (25.6)
College degree or greater	166 (27.0)	20 (25.6)
Unknown or not reported	63 (10.3)	10 (12.8)
Clinical characteristics		
Culture-positive Sepsis, n (%)	37 (6.0)	8 (10.3)
Severe NEC, n (%)	27 (4.4)	8 (10.3)
Grade III-IV IVH, n (%)	46 (7.5)	32 (41.0)
Severe BPD, n (%)	386 (62.9)	61 (78.2)
Received treatment with Epo, n (%)	45 (7.3)	14 (17.9)
Lowest Hematocrit over entire hospital course (%), mean (SD)	27.2 (5.1)	25.3 (3.8)
Lowest Ferritin over entire hospital course, mean (SD)	137 (139)	201 (191)
Transfusions and iron		
Infants who received prbc transfusion during NICU hospitalization, n (%)	491 (80.0)	73 (93.6)
Total transfusion volume in transfused infants (mL), mean (SD)	73 (74)	126 (97)
Infants who received oral iron, n (%)	613 (99.8)	77 (98.7)
Total oral iron dose (mg) in exposed, mean (SD)	641 (426)	663 (420)
Prolonged medication exposure		
>14 Days Dexamethasone, n (%)	44 (7.2)	13 (16.7)
>7 Days Opioids/Benzodiazepines, n (%)	297 (48.4)	65 (83.3)
Follow-up characteristics		
Infants who completed BSID subscale, n (%)		
Cognitive	614 (100)	78 (100)
Motor	603 (98.2)	77 (98.7)
Language	600 (97.7)	77 (98.7)
Mean (SD) BSID III Scores		
Cognitive	94.1 (13.2)	66.8 (12.4)
Motor	94.0 (12.9)	59.2 (11.0)
Language	91.0 (15.6)	65.3 (15.2)

Table 1: **Maternal and infant characteristics separated by severe neurodevelopmental impairment (NDI) at follow-up.**

model, the partial R^2 for each predictor was extracted using the rsq library in R, and the partial R^2 and -log10 (p-value) for each predictor was plotted in Prism version 9 (GraphPad software). The partial R^2 was used to infer the percent variance in the BSID-III subscale explained by each individual variable. For both the network analysis and GEE models, only infants who had data available for the BSID-III subscale of interest were included.

Outcome prediction using BART

To determine whether we could use PENUT data to reliably predict long-term neurodevelopmental outcomes in EP infants, we tested three predictive algorithms of increasing complexity: 1) using the five variables from the NICHD Extremely Preterm Birth Outcomes Tool, 2) using the 21 variables described in the network analysis section above, and 3) a hypothesis-free approach using all the potential variables available in the PENUT database in addition to the 21 network variables. Prediction models were constructed for cognitive, motor, and language BSID-III components as continuous variables as well as for severe NDI as a dichotomous outcome as defined by the original PENUT Trial.[15]

We performed hypothesis-free variable selection using the bartMachine package in R.[21] The predictor pool consisted of 133 variables including 57 baseline demographic and clinical features present in the first 24 h after birth, 55 clinical measurement variables that were measured once per subject throughout the child's hospital stay, 11 types of medication data, and 19 types of laboratory measurement data, along with GA (in days) and treatment assignment (placebo/Epo). The full list of variables is provided in Supplemental Table S1, including degree of missingness. Of these potential predictors, 79 had complete observations and 54 predictors exhibited missing values (maximum 55% missingness). Missingness was documented in two ways. For certain categorical and binary predictors, the documenting physician listed the information as "unknown" or "not reported". Other variables, including continuous predictors, were completely missing in the dataset. For the algorithm using the five NICHD variables, no data were missing. For the 21 pre-specified network predictors and 133 potential predictors screened by BART, categorical/binary variables documented as "unknown" or "not reported" were assigned the most common category (e.g., English for maternal language) or were assumed to not be present (e.g., "no" for yes/no binary predictors). Completely missing categorical or continuous variables were imputed using the mode or median value, respectively. We used this approach as it would be most likely to translate to a future prediction scenario where missing binary predictors would be assumed to not be present, categorical predictors would default to the most common category, and continuous predictors would default to the median value.

Potential predictors for each outcome were selected by BART using 10-fold cross-validation by permutation, with n = 100 permutation samples, and n = 50 trees for prediction in the held-out sample. The goal of variable selection was to identify a parsimonious prediction model, therefore n = 20 trees were used per permutation to force predictors to compete for inclusion in the model. We selected predictors whose variable inclusion proportion exceeded its local null threshold, which was defined as the 95th percentile of its permutation distribution. After variable selection we fit BART models using selected predictors for each outcome on the entire dataset to generate pre-validated predictions using 10-fold cross-validation. In practice, this means that we split the data into 90% training and 10% validation subsets a total of 10 times. This allows each infant to be included in a single validation fold as well as 9 training folds. As a result, we can achieve a prediction for each infant without overfitting to the data while maintaining predictive accuracy.[22] We followed the recommendations by Chipman et al.,[21] and the default number of 250 burn-in Gibbs samples and 1000 post-burn-in samples. We also defaulted all predictors to be equally important *a priori*.

We plotted true BSID-III scores against predicted scores for all subjects for each of the three levels of algorithm complexity to evaluate prediction error defined as the mean squared error (MSE). Locally estimated scatterplot smoothing (LOESS) curves were used in calibration plots to show how well a given average prediction compared to the actual observed values across the entire range of BSID-III scores. We further plotted the partial dependence (PD) of a subset of predictors for each score to examine the effect of changing a predictor after controlling for other predictors. The PD function of a predictor[23] gives the average value of the outcome, with 95% credible interval, showing how the average prediction changes across values of a predictor while other predictors remain the same. For improved visualization, PD objects from bartMachine were converted into ggplot objects using the pdplotGG function (available from https://github.com/CHEST-Lab/BART_Covid-19/blob/master/pdPlotGG.R). To allow for more interpretable PD plots for important categorical predictors, models were re-run using these variables as continuous predictors to generate PD plots only. For dichotomous severe NDI prediction, receiver operator characteristic (ROC) curves were plotted in addition to calibration plots for each model to examine the predicted probability compared to observed average probability of severe NDI. For each NDI prediction model, a predicted probability cut-off for NDI was selected that maximized the sum of sensitivity and specificity using the cutpointr library in R. This cut-off was used to determine the sensitivity and specificity of the models for predicting severe NDI.

Sensitivity analyses were performed to ensure that our approach to missingness did not affect our ability to predict outcomes. In these analyses, missing data were assumed to be not missing at random. As such, missingness was included as an attribute in variable selection and prediction to potentially improve predictive accuracy. Variables documented as "unknown" or "not reported" for categorical or binary outcomes were included as a separate category. Completely missing data were left as "NA" in the dataset and BART was instructed to use missingness as an attribute, as suggested by Twala, Jones, and Hand.[22] To ensure that the wide follow-up window did not affect our results, a second sensitivity analysis was performed including only the 625 infants who were assessed in the pre-specified PENUT follow-up window of 22–26 months CA.[15]

Role of the funding source

The funder had no role in the study design, data collection, data analysis, finding interpretation, or the writing of the manuscript. TRW, BAC, and SL had access to the data. All authors took the decision to submit the study results for publication.

Results

Maternal and neonatal data at the time of enrollment, clinical information derived from their hospital course, and BSID-III subscale outcomes are displayed in Table 1, separated by NDI status.

Network predictors directly associated with BSID component scores

Fig. 1A shows the graphical network of BSID-III cognitive score using the 21 pre-defined variables expected to be associated with long-term outcome. Strong relationships between many of the predictors are evident, with fewer directly influencing the BSID-III score. After considering how all the variables were related, maternal education had a strong positive association with cognition, while male sex, severe IVH (grade 3–4), and severe BPD were negatively associated. In a fully adjusted GEE model, maternal education, male sex, severe IVH, SGA, 5 min Apgar <5, and total oral iron prior to 36 weeks PMA were significantly associated with cognitive score. Fig. 1B shows the percent variance of cognitive score explained by each of the 21 predictors (partial R-squared values), with maternal education, severe IVH, male sex, SGA, 5 min Apgar <5, and oral iron up to 36 weeks PMA each predicting 2.5%, 2.2%, 1.9%, 1.0%, 0.9%, and 0.9% of the variance in cognitive score, respectively.

Fig. 2A shows the network analysis for the motor component of the BSID-III. After considering how all the variables were related in the network, severe IVH, sick appearance at birth, severe BPD, SGA, total transfusion volume, 5 min Apgar score <5, >14 days exposure to dexamethasone, Epo treatment, and male sex were all negatively associated with motor score, with a positive association between maternal education and motor score. In a fully adjusted GEE model, severe IVH, sick appearance at birth, severe BPD, SGA, total transfusion volume, 5 min Apgar score <5, >14 days exposure to dexamethasone, Epo treatment, and male sex treatment were significantly associated with motor score, with each predicting 4.7%, 1.7%, 1.6%, 1.3%, 1.2%, 1.2%, 0.9%, 0.8%, and 0.6% of the variance in motor score, respectively (Fig. 2B).

Fig. 3A shows the network analysis for BSID-III language score. There is a strong positive relationship between maternal education and language scores, with additional positive association between lowest hematocrit and language score. Variables significantly negatively associated with the language score were male sex, Epo treatment, SGA, total transfusion volume, 5 min Apgar <5. In a fully adjusted GEE model, maternal education, male sex, Epo treatment, SGA, and 5 min Apgar <5 were significantly associated with language, with each predicting 4.5%, 1.8%, 1.5%, 0.9%, and 0.6% of the variance in language score, respectively (Fig. 3B).

Important interactions/relationships between predictors

Graphical network analyses allow us to visualize important relationships between patient clinical characteristics and treatments (Figs. 1A–3A). For example,

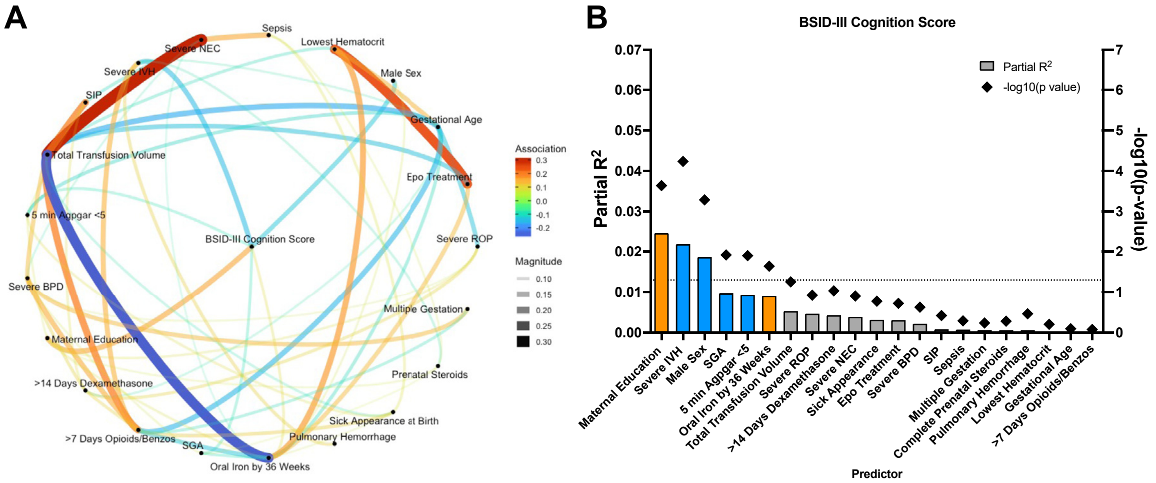

Fig. 1: **Network analysis of BSID-III Cognition Score.** (A) Graphical network analyses of BSID-III cognitive score using the 21 pre-defined variables from the PENUT dataset expected to be associated with long-term outcome. Line (edge) thickness between variables (nodes) shows the strength of association, and the color shows the directionality; green-blue lines depict a negative association, orange-red lines depict a positive association. After considering how all the variables were related, maternal education had a strong positive association with cognition, while male sex, severe IVH (grade 3–4), and severe BPD were negatively associated. In a fully adjusted GEE model, maternal education, male sex, severe IVH, SGA, 5 min Apgar <5, and total oral iron prior to 36 weeks PMA were significantly associated with cognitive score. In a fully adjusted GEE model (B), maternal education, severe IVH, male sex, SGA, 5 min Apgar <5, and oral iron up to 36 weeks PMA were significantly associated with cognitive score. The left y-axis (bars) depicts partial R^2 values in order of size of effect. For variables significantly associated with outcome, orange bars depict a positive association with cognition score and blue bars depict a negative association. Maternal education, severe IVH, male sex, SGA, 5 min Apgar <5, and oral iron up to 36 weeks PMA each predicted 2.5%, 2.2%, 1.9%, 1.0%, 0.9%, and 0.9% of the variance in cognitive score, respectively. The right y-axis (diamonds) depicts -log10 (p-value), with the dotted line at p = 0.05. As such, diamonds above the dotted horizontal line indicate p-values <0.05.

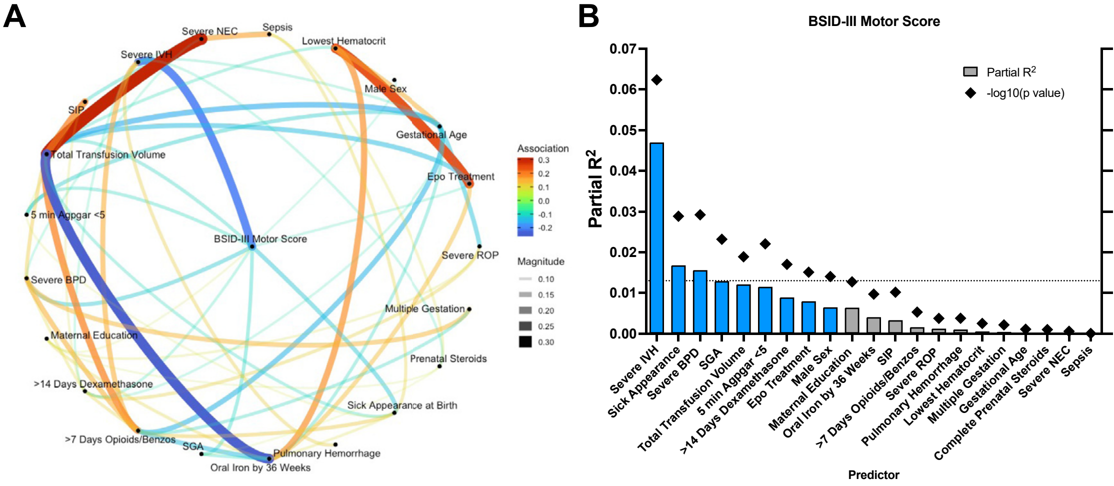

Fig. 2: **Network analysis of BSID-III Motor Score.** (A) Graphical network analyses of BSID-III motor score using the 21 pre-defined variables from the PENUT dataset expected to be associated with long-term outcome. Line (edge) thickness between variables (nodes) shows the strength of association, and the color shows the directionality; green-blue lines depict a negative association, orange-red lines depict a positive association. After considering how all the variables were related in the network, severe IVH, sick appearance at birth, severe BPD, SGA, total transfusion volume, 5 min Apgar score <5, >14 days exposure to dexamethasone, Epo treatment, and male sex were all negatively associated with motor score. Greater maternal education was associated with higher motor score. In a fully adjusted GEE model (B), severe IVH, sick appearance at birth, severe BPD, SGA, total transfusion volume, 5 min Apgar score <5, >14 days exposure to dexamethasone, Epo treatment, and male sex treatment were significantly associated with motor score. The left y-axis (bars) depicts partial R^2 values in order of size of effect. For variables significantly associated with outcome, orange bars depict a positive association with motor score and blue bars depict a negative association. Severe IVH, sick appearance at birth, severe BPD, SGA, total transfusion volume, 5 min Apgar score <5, >14 days exposure to dexamethasone, Epo treatment, and male sex treatment each predicted 4.7%, 1.7%, 1.6%, 1.3%, 1.2%, 1.2%, 0.9%, 0.8%, and 0.6% of the variance in motor score, respectively. The right y-axis (diamonds) depicts -log10 (p-value), with the dotted line at p = 0.05. As such, diamonds above the dotted horizontal line indicate p-values <0.05.

across all network analyses Epo treatment is significantly negatively associated with transfusion volume while being positively associated with lowest hematocrit. In order of the magnitude of the association, total transfusion volume is also positively associated with severe NEC, SIP, prolonged exposure to opioids and benzodiazepines, and severe BPD, while being negatively associated with oral iron administration. Sick appearance at birth is positively associated with culture-positive sepsis during hospitalization and >7 days of opioids/benzodiazepines. Higher GA is also associated with lower likelihood of having a 5 min Apgar score <5, lower likelihood of being exposed to >7 days of opioids/benzodiazepines, a higher lowest hematocrit, lower total transfusion volume, and lower incidence of ROP. However, GA was not independently associated with any of the Bayley subscales.

BSID-III prediction models

Fig. 4 shows true cognitive, motor, and language component BSID-III scores (Y axis) plotted against the predicted scores (X axis) using A) NICHD Extremely Preterm Birth Outcomes Tool variables, B) the 21 pre-selected variables, and C) predictors selected using BART in a hypothesis-free manner. As prediction improves, the dots align with the diagonal line, which indicates when the predicted score is the same as the actual score. We used mean squared error (MSE) to assess how far on average a prediction is from the actual value, to assess the accuracy of each model. The square root of the MSE gives us an estimate of how many BSID points on average the prediction is away from the true score.

NICHD outcome prediction tool variables

Using the five NICHD outcome prediction tool variables, the MSE of the resulting prediction of BSID-III subscale scores was 239 for cognitive, 273 for motor, and 301 for language score (Fig. 4A). The square root of the MSE associated with these predictions was 16–17 BSID-III points. This means that for any given prediction, the true individual value will on average lie 16–17 points above or below the prediction. As a result of this large margin of error, only 4.1%, 4.1%, and 3.3% of the variance in the cognitive, motor, and language scores, respectively, was explained by this prediction model.

Pre-selected clinical variables

Including the 21 pre-selected predictors to the GA model resulted in a 25–45 point reduction in MSE. Resulting MSEs were 213, 229, and 273 with 14.1%,

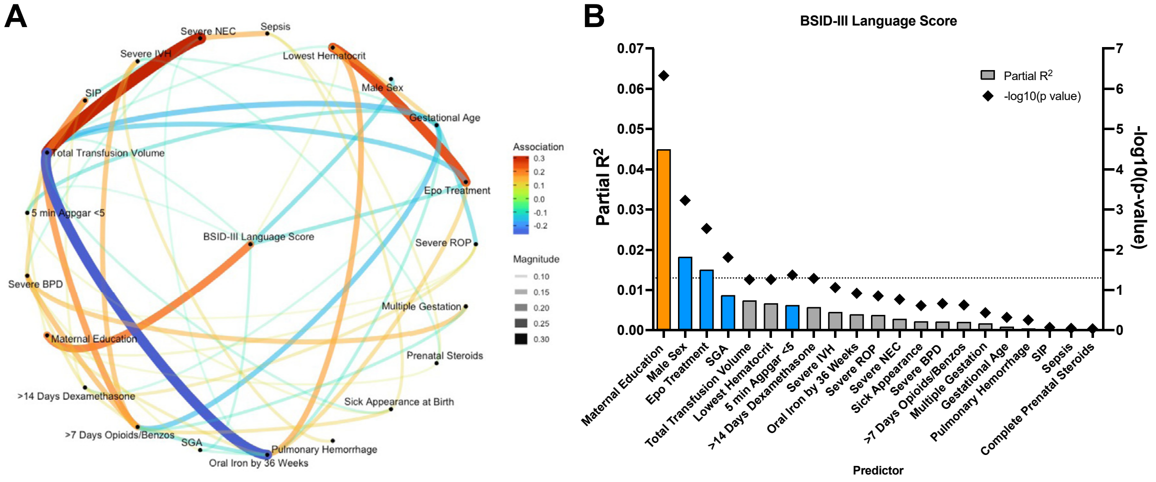

***Fig. 3:* Network analysis of BSID-III Language Score.** (A) Graphical network analyses of BSID-III language score using the 21 pre-defined variables from the PENUT dataset expected to be associated with long-term outcome. Line (edge) thickness between variables (nodes) shows the strength of association, and the color shows the directionality; green-blue lines depict a negative association, orange-red lines depict a positive association. After considering how all the variables were related in the network, maternal education and lowest hematocrit were positively correlated with language score Variables significantly negatively associated with the language score were male sex, Epo treatment, SGA, total transfusion volume, 5 min Apgar <5. In a fully adjusted GEE model (B), maternal education, male sex, Epo treatment, SGA, and 5 min Apgar <5 were significantly associated with language score. The left y-axis (bars) depicts partial R^2 values in order of size of effect. For variables significantly associated with outcome, orange bars depict a positive association with language score and blue bars depict a negative association. Maternal education, male sex, Epo treatment, SGA, and 5 min Apgar <5 each predicted 4.5%, 1.8%, 1.5%, 0.9%, and 0.6% of the variance in language score, respectively. The right y-axis (diamonds) depicts -log10 (p-value), with the dotted line at p = 0.05. As such, diamonds above the dotted horizontal line indicate p-values <0.05.

19.6%, and 12.1% of variance explained for cognitive, motor, and language scores, respectively (Fig. 4B). The improvement in MSE using these variables compared to the NICHD variable prediction corresponds to reducing the average error in the prediction by 1–2 BSID-III points; the square root of the MSE associated with these predictions was 15–16 points, or approximately 1 standard deviation (SD) in each of the BSID-III subscale scores.

BART outcome predictions – cross-validated BART variables
The final variables selected for inclusion in each of the prediction models are shown in Supplemental Table S1. Feeding status at discharge, maternal race, diagnosed hydrocephalus, treatment with vasopressors, and respiratory support after birth were included as variables in all BSID-III subscale models. Compared to using the 21 pre-selected variables, BART-selected variables produced predictions with a further 15–45 point reduction in MSE: 199, 195, and 226 with 19.8%, 31.0%, and 26.8% of variance explained for cognitive, motor, and language score, respectively (Fig. 4C). Though the percent of variance explained improved by an absolute 5–15%, the resulting square root of the MSE remained approximately 1 SD (14–15 BSID-III subscale points). LOESS curves of the average predicted versus actual score suggested that the BART variable models using all potential predictor variables showed a particular improvement in accuracy at lower scores (<75 points), particularly for language outcomes.

Importance of individual prediction variables

21 preselected variables
To determine which variables appeared to be most important for predicting each outcome, variable importance graphs were constructed which show how frequently each variable was included in a decision tree (Fig. S1). PD plots for the top six of the 21 preselected variables used to predict cognitive, motor, and language scores are shown in Supplemental Figs. S2–S4.

For cognitive score the two predictors most frequently included in trees were total transfusion volume and maternal education (Fig. S1A). Compared to infants who received <100 mL total transfusion volume, a total transfusion volume of 200 mL was associated with an estimated partial effect of −7 cognitive points (Fig. 5A). Compared to those with an education at high school level or less, having a bachelor's degree or greater was associated with an estimated partial effect of approximately +5 cognitive points (Fig. 5B). Having a severe IVH was also associated with a partial effect of around −6 cognition points (Fig. 5C).

For motor score, the two predictors most frequently included in trees were total transfusion volume and severe IVH (Fig. S1B). Compared to infants who

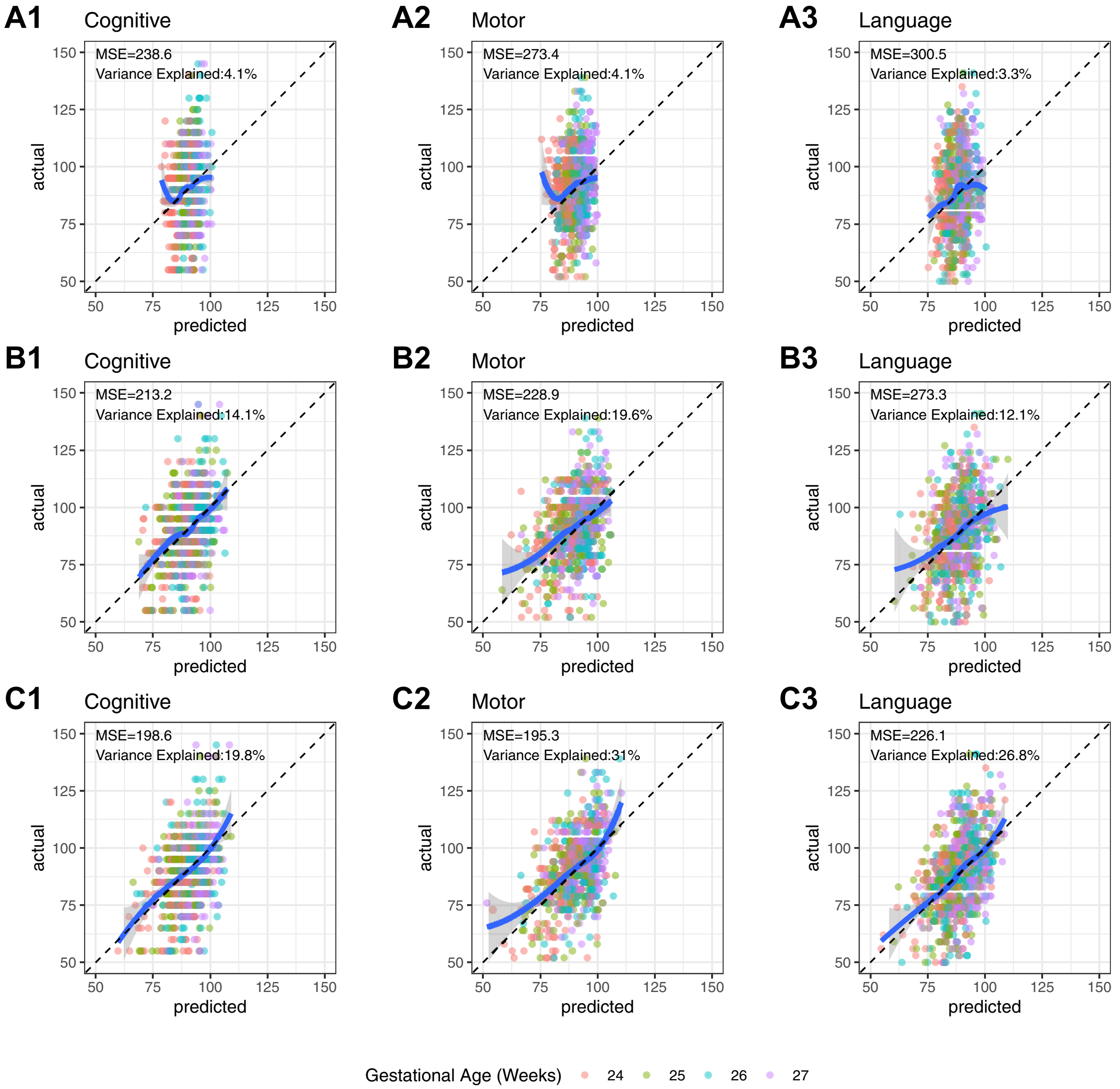

Fig. 4: **Calibration plots of pre-validated predictions.** Predicted (x-axis) versus actual (y-axis) BSID-III subscale scores based on revalidated predictions generated from 10-fold cross-validation using A) NICHD Extremely Preterm Birth Outcomes Tool variables, B) the 21 pre-selected variables clinically associated with long-term outcomes, and C) cross-validated predictors selected using BART in a hypothesis-free manner for each of the BSID-III components (Cognitive - 1, Motor - 2, Language - 3). Predictions for each infant are colored based on completed weeks of gestation at birth. A LOESS curve (blue, with 95% CI in grey) shows the average observed outcome across the range of predictions. As prediction improves, the dots approach alignment with the diagonal line, which indicates when the predicted score is the same as the actual score. We used MSE to assess how far on average a prediction is from the actual value. The square root of the MSE gives us an estimate of how many BSID points on average the prediction is away from the true score. MSE: Mean squared error.

received <100 mL total transfusion volume, a total transfusion volume of 200 mL was associated with an estimated partial effect of −10 motor points (Fig. 5D). Compared to those with an education at high school level or less, having a bachelor's degree or greater was associated with an estimated partial effect of approximately +3 motor points (Fig. 5E). Having a severe IVH was associated with a partial effect of around −10 motor points (Fig. 5F) and being exposed to >14 days of dexamethasone was associated with a partial effect of around −4 motor points (Fig. 5G).

For language score, the two predictors most frequently included in trees were maternal education and transfusion volume (Fig. S1C). Compared to infants who received <100 mL total transfusion volume, a total transfusion volume of 200 mL was associated with an

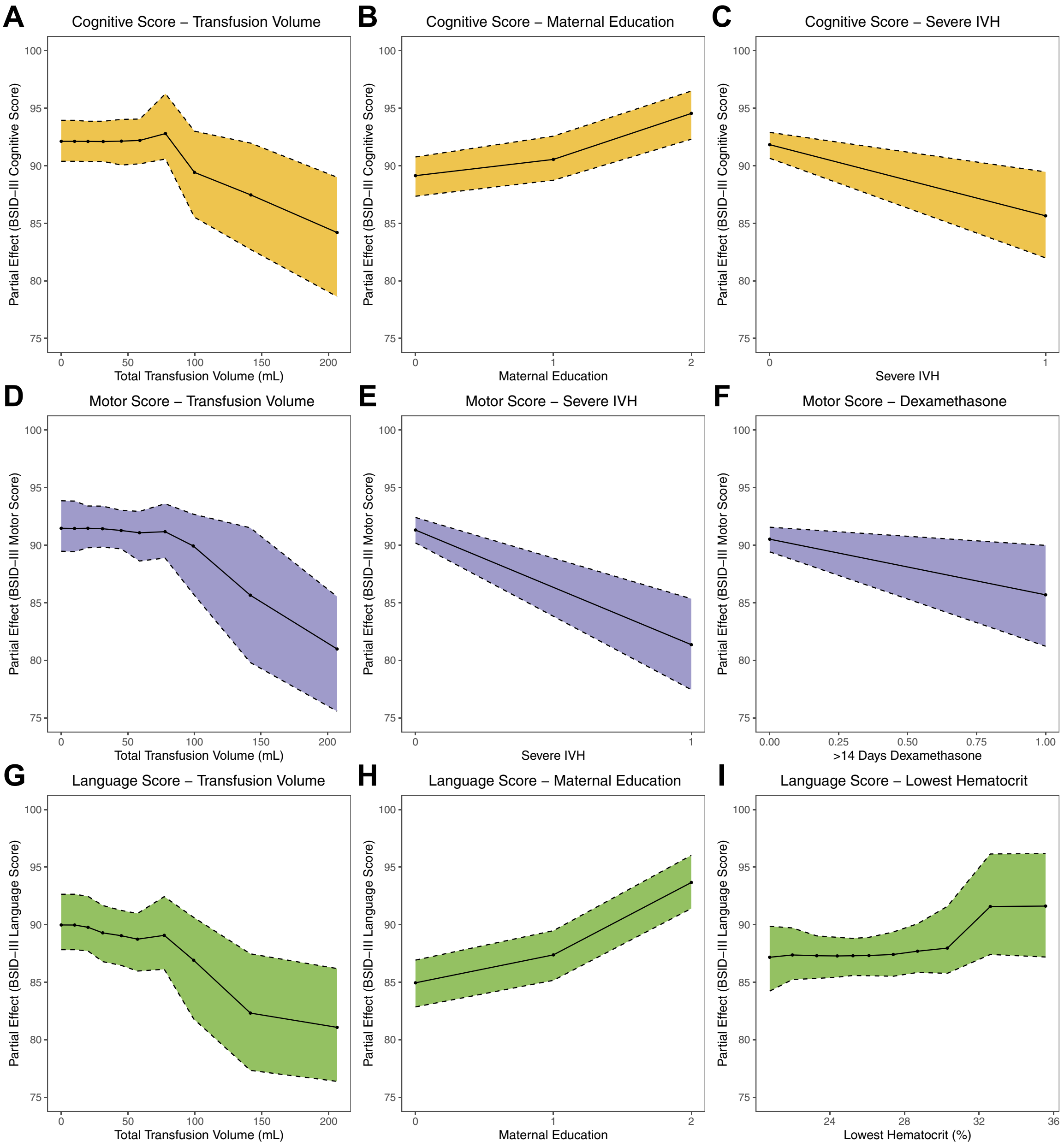

Fig. 5: **Partial dependence plots for BSID-III subscales using 21 network variables.** Selected partial dependence plots showing the most impactful variables associated with predicting cognitive (A–C), motor (D–F), and language (G–I) scores from the 21 network variables. Y-axis depicts partial effect of changing the variable on the x-axis while all other variables are held constant. Black lines depict average predicted effect, with colored shading indicating the 95% credible intervals. Each dot represents either the position of each decile (continuous variables) or discrete category (categorical and binary variables). Maternal education was classified as a three-level variable: 1 - high school or less, 2 - some college, or 3 - Bachelor's degree or higher.

estimated partial effect of −9 language points (Fig. 5G). Compared to those with an education at high school level or less, having a bachelor's degree or greater was associated with an estimated partial effect of approximately +8 language points (Fig. 5H). Compared to infants who had a lowest hematocrit <30%, having a lowest hematocrit >32% was associated with a partial effect of around +5 language points (Fig. 5I).

BART-selected variables

Variable importance for BSID-III predictions using BART-selected variables is shown in Fig. S5. PD plots for the top six most included BART variables used to predict cognitive, motor, and language scores are shown in Supplemental Figs. S6–S8.

For cognitive score, total transfusion volume, 1 min Apgar score, and maternal education were the three variables included in the greatest proportion of trees (Fig. S2A). However, 1 min Apgar score showed no clear pattern of partial effect (Fig. S6B). Compared to infants who received <100 mL total transfusion volume, a total transfusion volume of 200 mL was associated with an estimated partial effect of −6 cognitive points (Fig. 6A). Increasing levels of maternal education showed a linear partial effect with a total effect of around +5 cognitive points in the highest (graduate degree) versus lowest levels (Fig. 6B). Male sex was associated with a partial effect of around −3 cognitive points (Fig. 6C).

For motor score, total transfusion volume and highest direct bilirubin were the two variables included in the greatest proportion of trees (Fig. S2B). Compared to infants who received <100 mL total transfusion

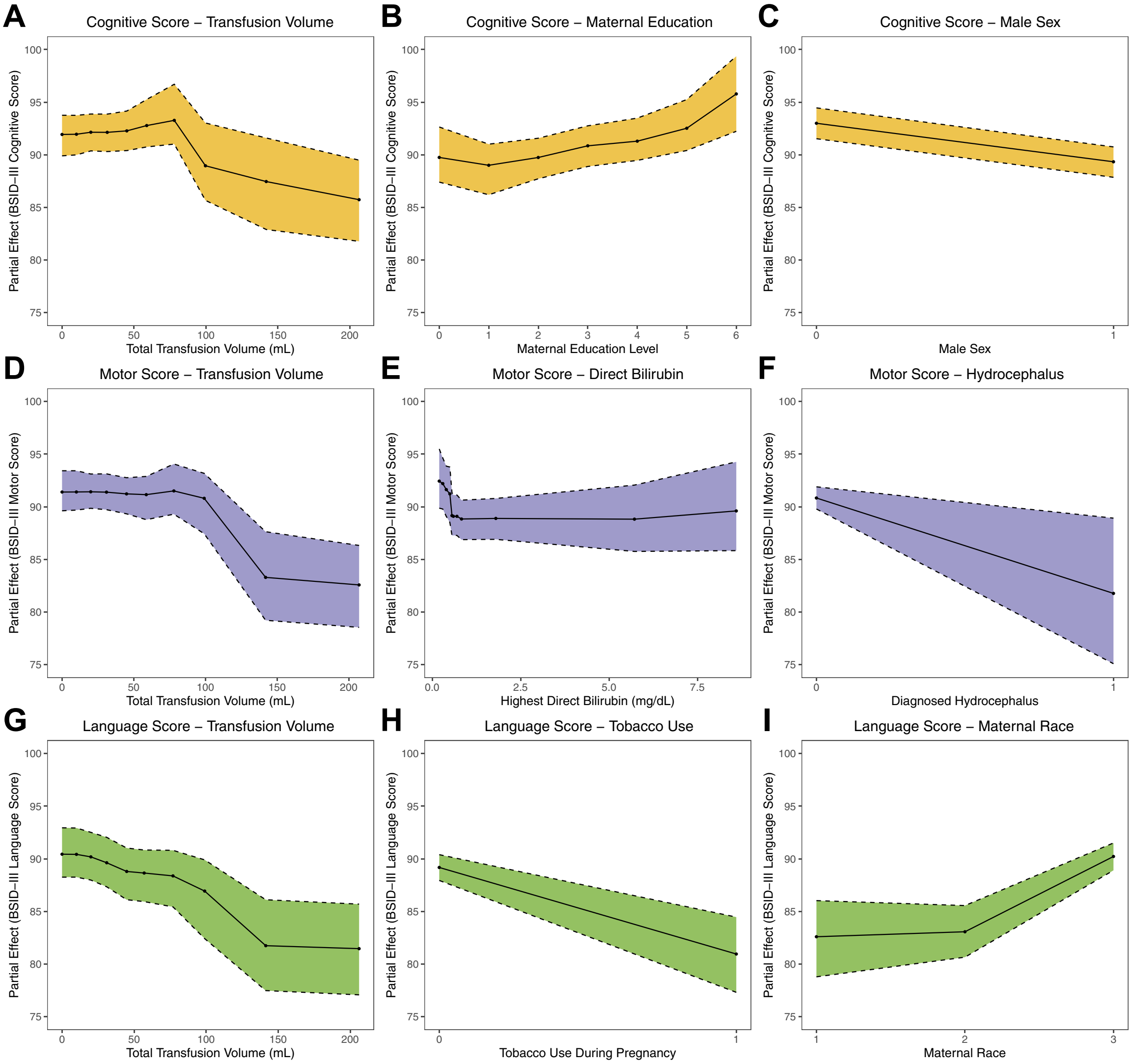

Fig. 6: **Partial dependence plots for BSID-III subscales using pre-validated BART variables.** Selected partial dependence plots showing the most impactful variables associated with predicting cognitive (A–C), motor (D–F), and language (G–I) scores using the pre-validated BART variables. Y-axis depicts partial effect of changing the variable on the x-axis while all other variables are held constant. Black lines depict average predicted effect, with colored shading indicating the 95% credible intervals. Each dot represents either the position of each decile (continuous variables) or discrete category (categorical and binary variables). Maternal education: 0 – never attended, kindergarten only, or unknown/unreported, 1 – less than high school, 2 - high school graduate/GED, 3 – some college (no degree), 4 – Associate's degree, 5 – Bachelor's degree, 6 – graduate degree. Race: 1 – Asian/Native American, Alaskan, or Pacific Islander, 2 – Black, 3 – white.

volume, a total transfusion volume of 200 mL was associated with an estimated partial effect of −9 motor points (Fig. 6D). Compared to infants with a highest direct bilirubin <0.5 mg/dL, a highest direct bilirubin >0.5 mg/dL was associated with a partial effect of around −4 motor points (Fig. 6E). Hydrocephalus was associated with a partial effect of around −8 motor points (Fig. 6F).

For language score, total transfusion volume, highest direct bilirubin, and male sex were the three variables that were included in the greatest proportion of trees (Fig. S2C). Compared to infants who received <100 mL total transfusion volume, a total transfusion volume of 200 mL was associated with an estimated partial effect of −8 language points (Fig. 6G). Compared to infants with a highest direct bilirubin <0.5 mg/dL, a highest direct bilirubin >0.5 mg/dL was associated with a partial effect of around −4 language points (Fig. S8B). Maternal tobacco use during pregnancy was associated with a partial effect of around −10 language points (Fig. 6H). Maternal race was also a strong predictor of language score – compared to infants whose mother did not identify as white, infants with a mother who identified as white were predicted to experience a partial effect of around +7 language points (Fig. 6I).

NDI predictions

Variable importance for predicting severe NDI using either the 21 network variables or BART-selected variables is shown in Fig. S9. Using the five NICHD outcome prediction variables resulted in an AUROC of 0.62 (95% CI 0.56–0.69). When selecting an NDI prediction probability cut-off that maximized the sum of sensitivity and specificity, the NICHD variable model displayed 71.8% sensitivity and 56.8% specificity for predicting severe NDI. The AUROC for predicting severe NDI improved to 0.77 (0.72–0.83) when using the 21 pre-selected network variables (Fig. 7), with an associated sensitivity of 70.5% and specificity of 75.6%. Of these 21 variables, severe IVH and total transfusion volume were the two variables that were included in the greatest proportion of trees, followed by SGA, male sex, >7 days exposure to opioids/benzodiazepines, and severe ROP (Fig. S9A, Fig. S10). As suggested by the improved accuracy of predicting lower BSID-III scores using BART-selected variables (Fig. 4), the AUROC for severe NDI prediction with BART variables was 0.87 (0.83–0.91; Fig. 7). The BART-selected variable model had a sensitivity of 84.6% and specificity of 72.3% for predicting severe NDI. Of all the BART variables, total transfusion volume was included in the greatest number of trees, with the next most important variables being male sex, hydrocephalus, and severe IVH (Fig. S9B, Fig. S11). Prediction using BART-selected variables also appeared to be well calibrated across the full range of predicted probability of severe NDI (Fig. S12).

Sensitivity analyses

In the first sensitivity analyses where missingness was coded as a separate category or used as an attribute during variable selection and prediction, the variables selected for the BART prediction models remained relatively similar (Table S3), with total transfusion volume remaining the variable included in the greatest number of trees for predicting all Bayley subscales as well as severe NDI (Fig. S13). Predictive performance did not meaningfully change compared to the primary analyses, with MSEs of 195, 200, and 227 with 21.4%, 29.4%, and 26.6% of variance explained for cognitive, motor, and language score, respectively (Fig. S14). The AUROC and 95% confidence interval for NDI prediction also remained entirely unchanged (0.87; 0.83–0.91, Fig. S17).

In the second sensitivity analysis that only included infants assessed in the 22–26-month window, the same predictors were largely selected by BART (Table S4). As with the other analyses, feeding status at discharge, electrographic seizures, total transfusion volume, infant sex, 1 min Apgar, hydrocephalus, white matter injury, and respiratory support after birth were selected as predictors for 3–4 of the outcomes. From the BART-selected variables, predictive performance was again similar. Final MSEs compared to the primary analysis were slightly higher - 208, 195, and 240 – with similar 20.6%, 33%, and 25% of variance explained for cognitive, motor, and language scores, respectively (Fig. S14). The AUROC and 95% confidence interval for NDI prediction also remained essentially unchanged (0.88; 0.84–0.91, Fig. S17).

Discussion

We present three models by which to predict long-term outcomes in EP infants, each with increasing complexity and increasing accuracy. Our baseline prediction model was based on the five variables used in the NICHD tool, which represents the best currently available outcome prediction tool for EP infants. The next set of prediction variables was selected based on strong biological premise as well as published literature suggesting potentially independent associations with long-term outcomes. The final set of prediction variables was selected using machine learning in a hypothesis-free manner from all the potential variables collected in PENUT. While this approach meaningfully improved predictive accuracy, the selected variables cannot necessarily be used to infer anything about biology – they are merely the variables for which the greatest signal is seen in the underlying data. It must be noted, however, that even though variables selected using BART resulted in notable improvements in MSE and variance explained by the model, the average individual prediction error remained large at around 1 SD either side of the true value. While dichotomous NDI

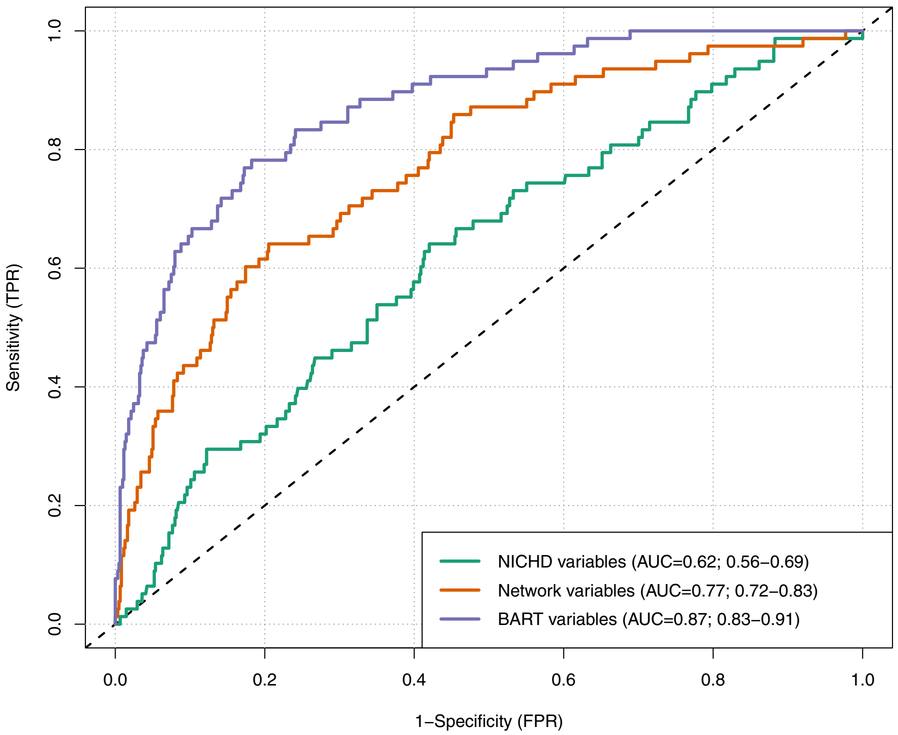

Fig. 7: **ROC curve for NDI prediction.** False positive rate (FPR) versus true positive rate (TPR) for NDI prediction based on 10-fold cross-validated predictions using the 5 NICHD outcome prediction variables, 21 pre-selected "network" variables clinically associated with long-term outcome, and cross-validated predictors selected using BART in a hypothesis-free manner. Legend shows AUROC with 95% confidence intervals.

prediction using BART variables was fairly accurate, and certain clinical variables such as transfusion exposure were meaningfully predictive of outcomes, these results underscore the fact that the field is still not able to predict complex long-term outcomes of infants born EP with high accuracy even when using rich datasets and advanced analytic methods.

When selecting a subset of variables from many potential predictors, several methods are available. While linear methods such as stepwise regression, Lasso, or elastic net regression are commonly used, they assume an underlying linear relationship and therefore rarely capture nonlinearities and interactions between predictors and outcome. In contrast, tree-based methods are better equipped to approximate complicated and relationships and interactions.[24] Therefore, we used BART for hypothesis-free predictor selection as well as for building the final prediction models which,[21] as expected, revealed a number of important non-linear relationships. This was particularly evident for transfusion volume, which was predictive of all BSID-III subscales as well as severe NDI, but the association was distinctly non-linear - decreases in scores were only seen as total volume increased above approximately 100 mL. This non-linearity may explain the relatively lower, though still significant, importance of transfusion volume in the network analyses, where a linear effect was assumed. A similar non-linear relationship between direct bilirubin levels and outcomes was seen, where a sharp decline in predicted motor and language score was seen around a level of 0.5 mg/dL, but no greater decrease above that.

While non-linear relationships appear to be important for maximizing predictive accuracy, we also employed linear graphical network analyses to allow us to more easily visualize important relationships between the various predictor variables. As expected, strong associations between Epo treatment, lowest hematocrit, and transfusion volume were seen, demonstrating that Epo increases erythropoiesis while decreasing transfusion volume. There were also significant associations between total transfusion volume and lowest hematocrit as well as severe NEC, SIP, BPD, and >7 days of opioids or benzodiazepines. The latter suggests that total transfusion may partly be a proxy for severity of illness, but we and others have shown an independent effect of transfusion volume and donor exposure on long-term outcomes and organ injury.[10,25,26] These negative associations may be related both to the detrimental effects of anemia as well as oxidative and inflammatory response to transfusions.[27–29] As expected, higher GA was associated with higher Apgar scores after birth, higher lowest hematocrit, lower total transfusion volume, less incidence of prolonged narcotic exposure, and lower incidence of ROP. However, GA was not independently associated with any of the Bayley scales, suggesting that the negative associations between GA and long-term outcomes in surviving EP infants may be largely mediated by postnatal events.

Different predictors were selected by BART to predict each component of the BSID-III; however, some factors were important for all three outcomes, particularly hydrocephalus and feeding status at discharge. Additional important predictors identified by BART included diuretic use (cognitive and motor scores), elevated direct bilirubin (motor and language scores), male sex (cognitive and language scores and NDI), elective Caesarean delivery (motor score and binary NDI outcome), and maternal tobacco use (all three Bayles subscale scores). Though they do not directly imply biological associations, these are variables that can either provide ground for future investigation or are already understood to have important associations with outcome. For example, infants with cholestasis may have experienced a longer duration of parenteral nutrition or may have been exposed to drugs or clinical illness that affect the liver.

Amongst both the 21 network variables and the BART-selected variables, maternal socioeconomic and environmental factors such as race, education, and tobacco use were frequently identified as important predictors of neurodevelopmental outcomes. Amongst the network variables, maternal education was strongly associated with higher cognitive and language scores, but less so for motor scores. In the BART-selected variables, maternal education was also selected as an important predictor of cognitive scores, with a linear benefit seen across the range of reported levels of education. In contrast, maternal tobacco use was associated with 10-point lower language scores. Having a mother who identified as white was also associated with a partial effect of around 7 language points. It is impossible to determine the mechanism of how these variables influence outcomes based on the prediction models we have used, but the significant impact of systemic factors is clear. For example, there are socioeconomic factors that differentially allow and encourage individuals of different backgrounds to pursue higher education; these include finances, the importance placed on education by family, and systemic inequity expressed as differential levels of oppression and access to educational opportunities throughout their life. However, we can speculate that a mother with higher education might talk to her child using a more complex vocabulary, thus improving the language abilities of their child. Similarly, tobacco exposure alone might influence outcomes, or more likely, the socioeconomic factors that endorse smoking might be most important. We also cannot rule out potentially discriminatory aspects of standardized testing, which have a history of being both racist and sexist.[30–32] Together, these factors reinforce the idea that the home and systemic environment remain the most significant determinants of long-term outcome in EP infants.[33]

For clinicians working to improve long-term outcomes of infants by improving care in the NICU, modifiable factors that were identified by both models included Epo treatment, oral iron supplementation up to 36 weeks PMA (but not later), lowest hematocrit, pRBC transfusions, sedation practices using opioids and benzodiazepines, and use of postnatal steroids. We speculate that the negative effect of Epo on language scores may be due to lower brain iron availability as iron was used to increase erythropoiesis, though this requires additional investigation. Whether or not erythropoiesis-stimulating agents are used, maintaining iron sufficiency and decreasing cumulative transfusion volume may improve outcomes. This study further suggests that limiting sedation and steroid use might be prudent.

Over the past few years, outcome prediction for EP infants has become an increasing focus for the field, both in terms of predicting death as well as long-term neurodevelopment in survivors. Crilly et al. recently summarized the field and identified dozens of prediction models and algorithms in the literature based on a varying array of variables and analytic tools.[14] However, the majority of models focused on dichotomous outcomes and assumed simple linear associations between variables and probability of outcome with no interaction between predictors. Most models were also not validated in other datasets and only used a single assessment of accuracy such as AUROC.

We have attempted to overcome some of these limitations. For instance, we used 10-fold cross-validation to avoid overfitting when generating predictions, and all models were assessed with multiple methods, including calibration - the degree to which the predicted outcome agrees with the average observed outcome across the entire range of predictions. BART also allows for nonlinear associations between predictors and outcome, and our network analyses consider how all the variables are associated with one-another before examining which variables are significantly associated with the outcome. By comparison, other machine-learning-based methods for prediction such as lasso or ridge regression require *a priori* determination of potential interactions and inflection points (knots) in the associations between continuous variables and outcomes through basis expansion, whereas our tree-based approach does not require additional engineering of the predictors to examine complex underlying relationships in the data. Though continuous outcome prediction remains a significant challenge, these advanced methods appear to be a step in the right direction, particularly for predicting NDI.

Our results appear to compare favorably to previous literature in this area. For instance, Broitman et al. predicted severe NDI in 2103 extremely low birth weight (<1000 g) infants using cranial ultrasound and clinical variables up to discharge. After training a model with 70% of the data, an AUROC of 0.68 was seen for NDI prediction in the remaining 30%.[34] Ambalavanan et al. also explored several prediction models in nearly 7000 infants from the Neonatal Research Network, with an AUROC for predicting NDI in survivors of 0.72 when using data available up to 36 weeks' PMA.[35] When summarizing the field, Crilly et al. note AUCs of up to 0.84 for predicting NDI; however, these results are from small cohorts and were likely to suffer from overfitting.[14] Therefore, we believe our NDI prediction accuracy with the BART-variable model (AUROC 0.87; 84.6% sensitivity, 72.3% specificity) is at least as good as what has been reported previously, but with a higher degree of external validity as predictions were cross-validated and we show good calibration across the range of predicted probabilities of NDI. By comparison to dichotomous outcome prediction, predicting continuous outcomes from complex multi-modal assessments such as the BSID-III subscales has not routinely been attempted. Though the final predictions from the BART-variable models retained significant error, predictions markedly improved when increasing the complexity of the data and allowing for complex interactions between variables. It is worth noting that different predictors were selected by BART for predicting the different outcomes. Only feeding status at discharge, diagnosed hydrocephalus, and total transfusion volume were selected for all BSID-III subscale models in addition to the severe NDI prediction. Improvements in prediction with those models will have been at least partly due to flexibility in variable selection, suggesting that the use of a fixed number or selection of predictor variables may constrain ability to predict a variety of complex outcomes. However, only 14–17 variables were required to make these predictions, with little evidence that more variables would be better. When even greater information was available to BART by using missingness as an attribute, a similar number of selected variables and predictive accuracy was seen, suggesting that we were likely to be nearing the limit of the predictive potential of the available dataset. This is unsurprising, as the majority of neurodevelopment relevant to the BSID-III subscales happens in the period after term-equivalent age, which is when EP infants have usually been discharged. Several studies have shown that the home environment and socioeconomic and parental factors have a dominant effect on the long-term outcomes of preterm infants,[36–39] and the PENUT cohort would be expected to be the same. Therefore, more detailed and longer-term assessments of the infant's post-discharge environment will be critical to improving prediction of neurodevelopmental outcomes.

This study has several limitations. As noted above, the variables and their associations with outcomes do not necessarily suggest causation or imply certain biological effects. For instance, a predictive signal from maternal SSRI use may be related to the impact of SSRIs themselves, maternal mental health conditions that are associated with SSRI use, societal factors that contribute to maternal mental health outcomes, some combination of these, or something else entirely. Our predictions are therefore limited by the nature of the data, which included maternal medications but not a full maternal mental health history. This study was also retrospective in nature, and some data for the selected variables were missing, creating the potential for selection bias. However, this was not an issue for the NICHD and selected BART variables (0% missingness for all), and minimally affected the 21 network variables. For missingness in the BART-selected variables, we chose to assume that binary predictors were not present, assigned the most common category to categorical predictors, and used the median value for missing continuous variables. This approach is most likely to translate to future predictions, for instance using an online calculator. Importantly, however, using missingness as an attribute did not improve predictions, which further supports our relatively conservative approach to missingness. We also recognize that, while neurodevelopmental assessment at 2 years CA is routine for long-term outcome determination in large neonatal clinical trials, NDI at 2 years does not necessarily predict NDI at a later age. For instance, a recent publication from the ELGAN (Extremely Low Gestational Age Newborn) study group found that nearly two-thirds of those classified as having moderate to severe NDI at 2 years had none to mild NDI at 10 years.[40] Furthermore, assessments of important outcomes such as autism, executive function, psychiatric symptoms (e.g., attention deficit, depression, and anxiety), and social and adaptive function are not feasible until the child is older.[40–44] This underscores the importance of long-term follow up studies. We must also acknowledge that children lost to follow-up tend to be from lower SES groups.[45] Another issue may be external validity with respect to the included population. Exclusion criteria for the PENUT trial included known life-threatening anomalies, chromosomal anomalies, disseminated intravascular coagulopathy, twin-to-twin transfusion, a hematocrit level above 65%, hydrops fetalis, or known congenital infection.[15] Therefore, these results may not be applicable to EP survivors who fit those criteria. More broadly, though the PENUT study was coordinated through academic centers, the study was performed in both academic and non-academic neonatal intensive care units in different regions of the United States, and our demographic data shows that the included infants were from fairly diverse backgrounds.[15,46] Our results were also similar when restricting the follow-up window from 20 to 33 months

CA to a more narrow 22–26 months (approximately 2-year) CA. Therefore, we believe that the results are likely to transfer to the majority of the EP population in the United States up to around three years CA, though this must be confirmed in future studies. Finally, we only focused on long-term outcome in survivors, with survival to assessment being a criterion for inclusion in the analysis.

In summary, we show that by using advanced analytic methods and a range of clinical and demographic predictor variables, most of which could be easily abstracted from medical records, we were able to meaningfully improve outcome prediction in EP survivors. Though the prediction models are not suitable for examining causation, we did identify a number of predictors that showed large partial effects on each BSID-III subscale, providing areas for future investigation as well as potential for improving clinical care if confirmed in other settings. For example, total transfusion volume was notably associated with all outcomes regardless of the variable selection strategy. However, though dichotomous NDI prediction using BART variables was fairly accurate, predicting complex continuous neurodevelopmental outcomes remains a challenge. Despite this, our approaches have significantly improved upon current outcome prediction tools, and we believe the principles used should form the basis of future work. This includes i) models that allow for non-linear and complex relationships between variables, ii) the potential for different variables to predict different outcomes, and iii) increasing external validity with methods such as cross-validation and employing multiple assessments of accuracy and calibration.[14] In order to further improve predictive accuracy in the future, these approaches should be applied to data that includes information about chronic health conditions and more in-depth details on the socio-economic and home environment. Some degree of usability must also be taken into account - while a complex model that requires dozens of predictors may have low usability or generalizability, a simple tool with few variables that has poor predictive ability also has limited utility. It is encouraging that most of the final variables selected by BART would be expected to be available for most EP infants and could easily be abstracted from the electronic medical record as part of an automated prediction tool. We believe that our results are an important step in the right direction, but they also reiterate the importance of the post-discharge environment on long-term outcome, highlighting the ongoing need for long-term follow-up of all EP infants.

Contributors

S.E.J., T.R.W., K.G., J.B.L., S.E.K., M.P-D., U.M., S.G., B.A.C., S.L., D.E.M., and P.J.H., conceived the manuscript. T.R.W. and B.A.C. verified the data. T.R.W. and S.L. performed the statistical analyses. T.R.W made the figures. S.E.J. and T.R.W. drafted the manuscript. All authors contributed to critical editing of the manuscript and approved the final draft. S.E.J. and P.J.H. oversaw the methodology and administration of the PENUT Trial and resulting investigations.

Data sharing statement

De-identified individual participant data from PENUT is available through the NINDS Data Archive: https://www.ninds.nih.gov/Current-Research/Research-Funded-NINDS/Clinical-Research/Archived-Clinical-Research-Datasets.

Declaration of interests

S.E.J. declares receiving royalties from editing Avery's Diseases of the Newborn. S.E.J., B.A.C., D.E.M., and P.J.H. received salary support as part of funding for the PENUT trial. All other authors declare that they have no relevant conflicts of interest.

Acknowledgements

The PENUT Trial was funded by the NINDS, U01NS077955 and U01NS077953.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2022.101782.

References

1 Martin JA, Hamilton BE, Osterman MJK, Driscoll AK. Births: final data for 2018. *Natl Vital Stat Rep*. 2019;68(13). National Center for Health Statistics.
2 Matthews TJ, MacDorman MF, Thoma ME. Infant mortality statistics from the 2013 period linked birth/infant death data set. National vital statistics reports : from the centers for disease control and prevention, national center for health statistics. *Nat Vital Stat Syst*. 2015;64(9):1–30.
3 Lee HC, Liu J, Profit J, et al. Survival without major morbidity among very low birth weight infants in California. *Pediatrics*. 2020;146(1).
4 Adams-Chapman I, Heyne RJ, DeMauro SB, et al. Neurodevelopmental impairment among extremely preterm infants in the neonatal Research network. *Pediatrics*. 2018;141(5).
5 Younge N, Goldstein RF, Bann CM, et al. Survival and neurodevelopmental outcomes among periviable infants. *N Engl J Med*. 2017;376(7):617–628.
6 Bell EF, Hintz SR, Hansen NI, et al. Mortality, in-hospital morbidity, care practices, and 2-year outcomes for extremely preterm infants in the US, 2013-2018. *JAMA*. 2022;327(3):248–263.
7 Cheong JLY, Olsen JE, Lee KJ, et al. Temporal trends in neurodevelopmental outcomes to 2 Years after extremely preterm birth. *JAMA Pediatr*. 2021;175(10):1035–1042.
8 Shah PS, Ye XY, Synnes A, et al. Prediction of survival without morbidity for infants born at under 33 weeks gestational age: a user-friendly graphical tool. *Arch Dis Child Fetal Neonatal*. 2012;97(2):F110-F115.
9 Bassler D, Stoll BJ, Schmidt B, et al. Using a count of neonatal morbidities to predict poor outcome in extremely low birth weight infants: added role of neonatal infection. *Pediatrics*. 2009;123(1):313–318.
10 Vu PT, Ohls RK, Mayock DE, et al. Transfusions and neurodevelopmental outcomes in extremely low gestation neonates enrolled in the PENUT Trial: a randomized clinical trial. *Pediatr Res*. 2021;90(1):109–116.
11 German KR, Vu PT, Comstock BA, et al. Enteral iron supplementation in extremely preterm infants and its positive correlation with neurodevelopment; post hoc analysis of the PENUT randomized controlled trial. *J Pediatr*. 2021;238:102-109.e8.
12 Puia-Dumitrescu M, Comstock BA, Li S, et al. Assessment of 2-year neurodevelopmental outcomes in extremely preterm infants receiving opioids and benzodiazepines. *JAMA Netw Open*. 2021;4(7):e2115998.
13 Puia-Dumitrescu M, Wood TR, Comstock BA, et al. Dexamethasone, prednisolone, and methylprednisolone use and 2-year neurodevelopmental outcomes in extremely preterm infants. *JAMA Netw Open*. 2022;5(3):e221947.

14 Crilly CJ, Haneuse S, Litt JS. Predicting the outcomes of preterm neonates beyond the neonatal intensive care unit: what are we missing? *Pediatr Res*. 2021;89(3):426–445.

15 Juul SE, Comstock BA, Wadhawan R, et al. A randomized trial of Erythropoietin for neuroprotection in preterm infants. *N Engl J Med*. 2020;382(3):233–243.

16 Williams DR, Rast P. Back to the basics: rethinking partial correlation network methodology. *Br J Math Stat Psychol*. 2020;73(2):187–212.

17 Bell MJ, Ternberg JL, Feigin RD, et al. Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. *Ann Surg*. 1978;187(1):1–7.

18 Law JB, Wood TR, Gogcu S, et al. Intracranial hemorrhage and 2-year neurodevelopmental outcomes in infants born extremely preterm. *J Pediatr*. 2021;238:124-134.e10.

19 R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical computing; 2019.

20 Liang KY, Zeger SL. Regression analysis for correlated data. *Annu Rev Public Health*. 1993;14:43–68.

21 Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266–298.

22 Twala BETH, Jones MC, Hand DJ. Good methods for coping with missing data in decision trees. *Pattern Recogn Lett*. 2008;29(7):950–956.

23 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001:1189–1232.

24 Bleich J, Kapelner A, George EI, Jensen ST. Variable selection for BART: an application to gene regulation. *Ann Appl Stat*. 2014:1750–1781.

25 Lust C, Vesoulis Z, Jackups Jr R, et al. Early red cell transfusion is associated with development of severe retinopathy of prematurity. *J Perinatol*. 2019;39(3):393–400.

26 Wang YC, Chan OW, Chiang MC, et al. Red blood cell transfusion and clinical outcomes in extremely low birth weight preterm infants. *Pediatr Neonatol*. 2017;58(3):216–222.

27 Patel RM, Knezevic A, Shenvi N, et al. Association of red blood cell transfusion, anemia, and necrotizing enterocolitis in very low-birth-weight infants. *JAMA*. 2016;315(9):889–897.

28 Arthur CM, Nalbant D, Feldman HA, et al. Anemia induces gut inflammation and injury in an animal model of preterm infants. *Transfusion*. 2019;59(4):1233–1245.

29 Hirano K, Morinobu T, Kim H, et al. Blood transfusion increases radical promoting non-transferrin bound iron in preterm infants. *Arch Dis Child Fetal Neonatal Ed*. 2001;84(3):F188–F193.

30 Fryer RGJ, Levitt SD. Testing for racial differences in the mental ability of young children. *Am Econ Rev*. 2013;103(2):981–1005.

31 Cooper RS. Race and IQ: molecular genetics as deus ex machina. *Am Psychol*. 2005;60(1):71–76.

32 Zoref L, Williams P. A look at content bias in IQ tests. *J Educ Measure*. 1980;17(4):313–322.

33 Benavente-Fernández I, Synnes A, Grunau RE, et al. Association of socioeconomic status and brain injury with neurodevelopmental outcomes of very preterm children. *JAMA Netw Open*. 2019;2(5):e192914.

34 Broitman E, Ambalavanan N, Higgins RD, et al. Clinical data predict neurodevelopmental outcome better than head ultrasound in extremely low birth weight infants. *J Pediatr*. 2007;151(5):500-5–505.e1-2.

35 Ambalavanan N, Carlo WA, Tyson JE, et al. Outcome trajectories in extremely preterm infants. *Pediatrics*. 2012;130(1):e115–e125.

36 Benavente-Fernández I, Synnes A, Grunau RE, et al. Association of socioeconomic status and brain injury with neurodevelopmental outcomes of very preterm children. *JAMA Netw Open*. 2019;2(5):e192914-e.

37 Joseph RM, O'Shea TM, Allred EN, et al. Maternal educational status at birth, maternal educational advancement, and neurocognitive outcomes at age 10 years among children born extremely preterm. *Pediatr Res*. 2018;83(4):767–777.

38 Wong HS, Edwards P. Nature or nurture: a systematic review of the effect of socio-economic status on the developmental and cognitive outcomes of children born preterm. *Matern Child Health J*. 2013;17(9):1689–1700.

39 Burnett AC, Cheong JLY, Doyle LW. Biological and social influences on the neurodevelopmental outcomes of preterm infants. *Clin Perinatol*. 2018;45(3):485–500.

40 Taylor GL, Joseph RM, Kuban KCK, et al. Changes in neurodevelopmental outcomes from age 2 to 10 years for children born extremely preterm. *Pediatrics*. 2021;147(5):e2020001040.

41 Msall ME, Buck GM, Rogers BT, Catanzaro NL. Kindergarten readiness after extreme prematurity. *Am J Dis Child*. 1992;146(11):1371–1375.

42 Wong HS, Santhakumaran S, Cowan FM, et al. Developmental assessments in preterm children: a meta-analysis. *Pediatrics*. 2016;138(2).

43 Johnson S, Fawke J, Hennessy E, et al. Neurodevelopmental disability through 11 years of age in children born before 26 weeks of gestation. *Pediatrics*. 2009;124(2):e249–e257.

44 O'Shea TM, Joseph RM, Allred EN, et al. Accuracy of the Bayley-II mental development index at 2 years as a predictor of cognitive impairment at school age among children born extremely preterm. *J Perinatol*. 2018;38(7):908–916.

45 Callanan C, Doyle L, Rickards A, et al. Children followed with difficulty: how do they differ? *J Paediatr Child Health*. 2001;37(2):152–156.

46 Ponnapakkam A, Carr NR, Comstock BA, et al. Factors associated with outpatient therapy utilization in extremely preterm infants. *Am J Perinatol*. 2021. https://doi.org/10.1055/a-1692-0544. Online ahead of print.