Title

Artificial Intelligence Outcome Prediction in Neonates with Encephalopathy (AI-OPiNE).

Journal

Radiology: Artificial Intelligence, 6(5)

Authors

Lew, Christopher
Calabrese, Evan
Chen, Joshua
et al.

Publication Date

2024-09-01

DOI

10.1148/ryai.240076

Peer reviewed

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Artificial Intelligence Outcome Prediction in Neonates with Encephalopathy (AI-OPiNE)

Christopher O. Lew, MD • Evan Calabrese, MD, PhD* • Joshua V. Chen, MD • Felicia Tang, BA • Gunvant Chaudhari, MD • Amanda Lee, MD • John Faro, MD • Sandra Juul, MD, PhD • Amit Mathur, MD • Robert C. McKinstry, MD, PhD • Jessica L. Wisnowski, PhD • Andreas Rauschecker, MD, PhD • Yvonne W. Wu, MD, MPH • Yi Li, MD*

From the Department of Radiology, Duke University Medical Center, 2301 Erwin Rd, Box 3808, Durham, NC 27710 (C.O.L., E.C, A.L., J.F.); Department of Radiology (J.V.C., F.T., G.C., A.R., Y.L.) and Weill Institute for Neurosciences (Y.W.W.), University of California San Francisco, San Francisco, Calif; Department of Pediatrics, University of Washington, Seattle, Wash (S.J.); Department of Pediatrics, Saint Louis University, St Louis, Mo (A.M.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (R.C.M.); and Children's Hospital Los Angeles, University of Southern California, Los Angeles, Calif (J.L.W.). Received February 5, 2024; revision requested March 11; revision received May 21; accepted June 18. **Address correspondence to** E.C. (email: *evan.calabrese@duke.edu*).

* C.O.L. and E.C. contributed equally to this work.

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

See also commentary by Rafful and Reis Teixeira in this issue.

Radiology: Artificial Intelligence 2024; 6(5):e240076 • https://doi.org/10.1148/ryai.240076 • Content codes: AI NR PD

Purpose: To develop a deep learning algorithm to predict 2-year neurodevelopmental outcomes in neonates with hypoxic-ischemic encephalopathy using MRI and basic clinical data.

Materials and Methods: In this study, MRI data of term neonates with encephalopathy in the High-dose Erythropoietin for Asphyxia and Encephalopathy (HEAL) trial (ClinicalTrials.gov: NCT02811263), who were enrolled from 17 institutions between January 25, 2017, and October 9, 2019, were retrospectively analyzed. The harmonized MRI protocol included T1-weighted, T2-weighted, and diffusion tensor imaging. Deep learning classifiers were trained to predict the primary outcome of the HEAL trial (death or any neurodevelopmental impairment at 2 years) using multisequence MRI and basic clinical variables, including sex and gestational age at birth. Model performance was evaluated on test sets comprising 10% of cases from 15 institutions (in-distribution test set, $n = 41$) and 10% of cases from two institutions (out-of-distribution test set, $n = 41$). Model performance in predicting additional secondary outcomes, including death alone, was also assessed.

Results: For the 414 neonates (mean gestational age, 39 weeks ± 1.4 [SD]; 232 male, 182 female), in the study cohort, 198 (48%) died or had any neurodevelopmental impairment at 2 years. The deep learning model achieved an area under the receiver operating characteristic curve (AUC) of 0.74 (95% CI: 0.60, 0.86) and 63% accuracy in the in-distribution test set and an AUC of 0.77 (95% CI: 0.63, 0.90) and 78% accuracy in the out-of-distribution test set. Performance was similar or better for predicting secondary outcomes.

Conclusion: Deep learning analysis of neonatal brain MRI yielded high performance for predicting 2-year neurodevelopmental outcomes.

Clinical trial registration no. NCT02811263

Supplemental material is available for this article.

Hypoxic-ischemic encephalopathy (HIE) caused by perinatal birth asphyxia is a major cause of death and long-term neurologic disability among infants worldwide, affecting more than 3 million neonates annually (1). Reduced cerebral blood flow from hypoxia in the perinatal period can lead to neuronal cell death and permanent neurologic injury. Although the initial diagnosis of neonatal encephalopathy is made on a clinical basis, neonatal brain MRI within the 1st week after injury plays an important role in determining the presence, location, and severity of injury and in counseling neurodevelopmental prognosis (2,3). Neonatal brain MRI can also serve as an early marker of outcome, allowing for early determination of the efficacy of novel neuroprotective treatments, many of which are currently under development (4).

To quantify brain injury for clinical trials and developmental prognostication, multiple semiquantitative scoring systems have been created to capture the injury severity and location. Neonatal brain imaging is challenging to interpret, and these scoring systems are subject to interrater variability, even among expert readers. In a 2014 study, three trained radiologists scored MR images of infants with HIE using a designated scoring system; analysis of agreement for scoring apparent diffusion coefficient sequences and T1- or T2-weighted images demonstrated a κ range from 0.27 to 0.66 and -0.11 to 0.44, respectively (5). These challenges with interpretation and consistency highlight the need to develop additional advanced image analysis methods to aid with standardizing interpretation and prognostication.

There have been few prior reports to date on the use of deep learning to study neonatal brain MRI (6), with novel efforts on the horizon (7). In this study, we developed a deep learning algorithm using neonatal brain MRI and

Abbreviations

AUC = area under the receiver operating characteristic curve, HEAL = High-dose Erythropoietin for Asphyxia and Encephalopathy, HIE = hypoxic-ischemic encephalopathy, NDI = neurodevelopmental impairment, OPiNE = Outcome Prediction in Neonates with Encephalopathy

Summary

In neonates with encephalopathy, a deep learning algorithm effectively identified patients who experienced death or neurodevelopmental impairment at 2 years using multisequence MRI and clinical data gathered 4–6 days after birth.

Key Points

- This was a secondary analysis of a deep learning study of 414 neonates from the prospective High-dose Erythropoietin for Asphyxia and Encephalopathy trial with multisequence MRI brain examinations and 2-year clinical follow-up.
- A multichannel image-based deep learning model predicted death or any neurodevelopmental impairment at 2 years with an area under the receiver operating characteristic curve (AUC) of 0.74 (95% CI: 0.60, 0.86) in the internal test set and an AUC of 0.77 (95% CI: 0.63, 0.90) in the external test set.
- Additional models were developed to predict secondary outcomes, including moderate and severe neurodevelopmental impairment and death, with AUCs ranging from 0.75 to 0.92 in the internal test set and 0.85 to 0.95 in the external test set.

Keywords

Convolutional Neural Network (CNN), Prognosis, Pediatrics, Brain, Brain Stem

perinatal clinical variables obtained as part of the harmonized, multisite High-dose Erythropoietin for Asphyxia and Encephalopathy (HEAL) trial to predict the primary outcome of death or neurodevelopmental impairment (NDI) at 2 years.

Materials and Methods

Study Sample

This was a Duke Health Institutional Review Board–approved, retrospective analysis of data from the HEAL trial (ClinicalTrials.gov: NCT02811263) (8–10). Participants were enrolled according to state and federal Health Insurance Portability and Accountability Act regulations, with written informed parental consent. The HEAL study prospectively enrolled 500 neonates (birth at 36 weeks or greater gestational age) with moderate to severe encephalopathy across 17 institutions within the United States between January 25, 2017, and October 9, 2019. Half of enrolled participants were randomized to receive human recombinant erythropoietin, a cytokine with neuroprotective and neuroregenerative effects in preclinical models of neonatal HIE (11), as an adjuvant therapy in addition to therapeutic hypothermia. The trial ultimately concluded no risk reduction of death or NDI with erythropoietin compared with placebo (9). Exclusion criteria for this secondary analysis of HEAL trial imaging, clinical, and outcome data were missing or incomplete MRI data (n = 70) and missing or incomplete clinical follow-up data (n = 16) (Fig 1). All image data were de-identified at each respective site using institutional review board–approved methods.

Participant Outcomes

Participants in the HEAL trial were evaluated at 24 months of age using the Bayley Scales of Infant Development III, a standardized neurologic examination (12), and the Gross Motor Function Classification System score (13). NDI was defined as any of the following: Gross Motor Function Classification System level of 1 or greater, 0, or 0.5 and cerebral palsy on neurologic examination, or Bayley Scales of Infant Development III cognitive score of less than 90 (± 0.67 [SD] below the mean) (9). The primary outcome in the HEAL trial was death or any NDI at 2 years of age (22–36 months), which we used as the primary end point for our current analysis. Secondary outcomes in our analysis included an ordinal stratification of no, mild, moderate, or severe NDI or death. This stratified severity of NDI was determined by the worst severity observed in either cognitive or motor outcome at 2 years (14). Cognitive outcome was measured by the Bayley Scales of Infant Development III, with severity of impairment defined as follows: normal, 90 or more; mild, 85–89; moderate, 70–84; and severe, less than 70. Motor outcome was defined by the presence of cerebral palsy (defined by a standardized neurologic examination) and by a modified Gross Motor Function Classification System.

Data Splits

Study participants were divided into a discovery set consisting of neonates proportionally sampled from 15 of 17 institutions (332 of 414 [80%] of total neonates), an in-distribution test set consisting of neonates proportionally sampled from the same 15 institutions (41 of 414 [10%] of total neonates), and an out-of-distribution test consisting of all neonates from the remaining two institutions (41 of 414 [10%] of total neonates). The out-of-distribution test set was chosen manually based on the following criteria: *(a)* all cases from the selected site(s) represented approximately 10% of the total dataset and *(b)* the prevalence of death or NDI in these cases approximated that of the total dataset (approximately 48%). All possible combinations of one or more sites were considered, and the final combination of two sites most closely matched the desired criteria. Discovery set participants were randomly divided into training (80%) and validation (20%), stratified by institution. Both in- and out-of-distribution test set participants were excluded from training and validation and were only used for final model assessment.

Image Data Preprocessing

De-identified Digital Imaging and Communications in Medicine data from the HEAL trial were converted to the Neuroimaging Informatics Technology Initiative format using dcm2niix (15). MRI data subsequently underwent standard image preprocessing, including brain extraction and coregistration with a custom deep learning method. T1- and T2-weighted images were then corrected for intensity nonuniformity using N4 bias correction and normalized to between 0 and 1 by bounding the minimum and maximum 0.1 percentile intensities (16). Diffusion trace and apparent diffusion coefficient images were calculated from diffusion tensor imaging data using FSL ver-

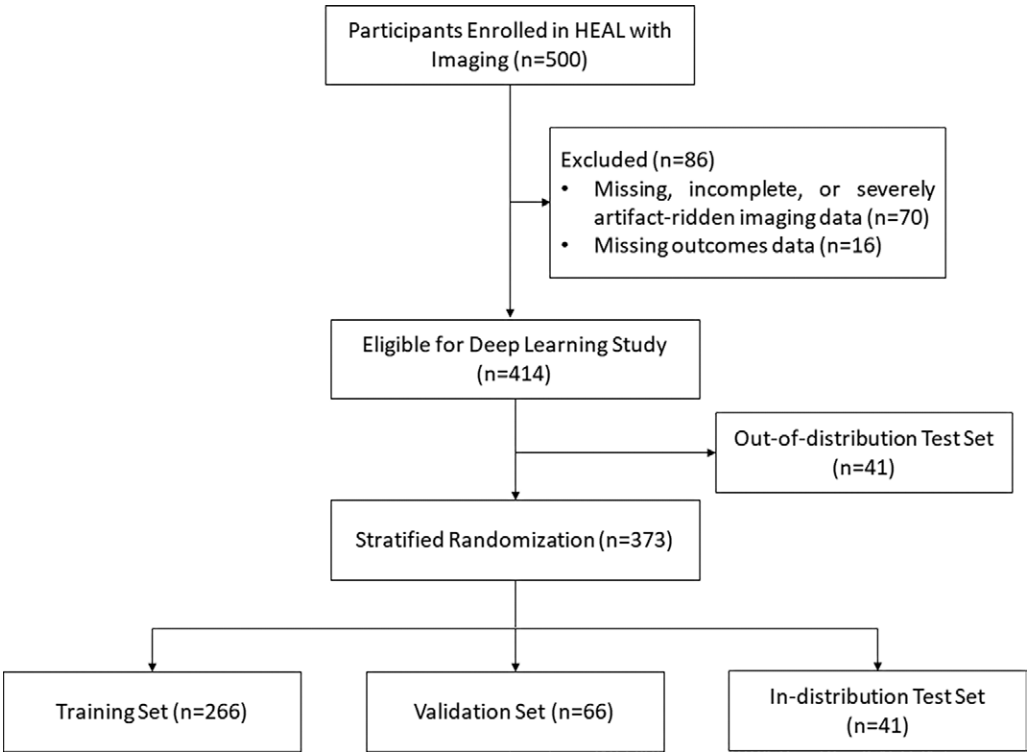

Figure 1: Flow diagram for participants included in each subset. Participants were enrolled as part of the High-dose Erythropoietin for Asphyxia and Encephalopathy (HEAL) study from 17 different institutions within the United States. The out-of-distribution test set contains all participants from two institutions. Participants from the remaining 15 institutions were randomly split into a training set, validation set, and in-distribution test set with an equal distribution of institutions in each subset.

sion 6.0.2 (FMRIB, *https://fsl.fmrib.ox.ac.uk/fsl/docs/#/*). Images were cropped to minimize empty voxels, yielding dimensions of 96 × 112 × 96 with 1.0-mm^3 isotropic voxel size. Representative participant MR images are shown in Figure 2, and an overview of data preprocessing is shown in Figure 3A.

Tabular Clinical Data

Tabular clinical data collected as part of the HEAL trial included sex, gestational age, and erythropoietin administration. Brain injury volume, defined as the number of voxels with apparent diffusion coefficient values of less than 800 × 10^{-6} mm^2/sec, was included as it has shown a correlation to poor outcomes in neonates with HIE (17). Using the discovery set, all tabular data were presented to the model in standardized forms, both by scaling the data between 0 and 1 and by using *z* score normalization. There were no missing tabular data.

Model Development

The Outcome Prediction in Neonates with Encephalopathy (OPiNE) model was developed as a convolutional neural network classifier incorporating multichannel MRI and tabular data to predict neurodevelopmental outcomes at 2 years. Multiple three-dimensional preprocessed MRI sequences were combined to input into the model as a multichannel (four-dimensional) array. Architecture was based on a standard convolutional neural network encoder followed by linear layers for classification. The binary cross-entropy loss function and AdamW optimizer were used.

Discovery set data were used to optimize model hyperparameters via a grid search of the following parameters: number of convolutional blocks [2, 3, 4, 5], number of linear layers [1, 2, 3], learning rate (3e-3, 3e-4, 3e-5), batch size (2, 4, 8), and total epochs (50, 100, 200). The final model architecture consisted of four convolutional blocks followed by two linear layers with dropout (30%) and a final sigmoid transform to yield predictions. The following hyperparameters were the result of the grid search in the OPiNE model: learning rate of 0.0003, batch size of 4, and total epochs of 100. Convolutional layers used Kaiming initialization (18). Tabular data were combined with image-derived features at the first linear layer.

After completing the grid search, the model was trained one final time using the training and validation data. An overview of the ensemble deep learning method is shown in Figure 3B. Similar models were developed using individual MRI channels alone and tabular data alone. Code repository can be found at *https://github.com/chris-lew/neonatal_HIE_outcome_prediction*.

Model Training

Model training was accomplished using a Linux Docker container on a desktop workstation with two NVIDIA RTX A6000 graphics processing units. MRI data were augmented during training by mirroring the images along different axes. For the primary outcome analysis, both the convolutional neural network and linear layers were trained. For secondary outcome analyses, we retrained the final linear layer to predict

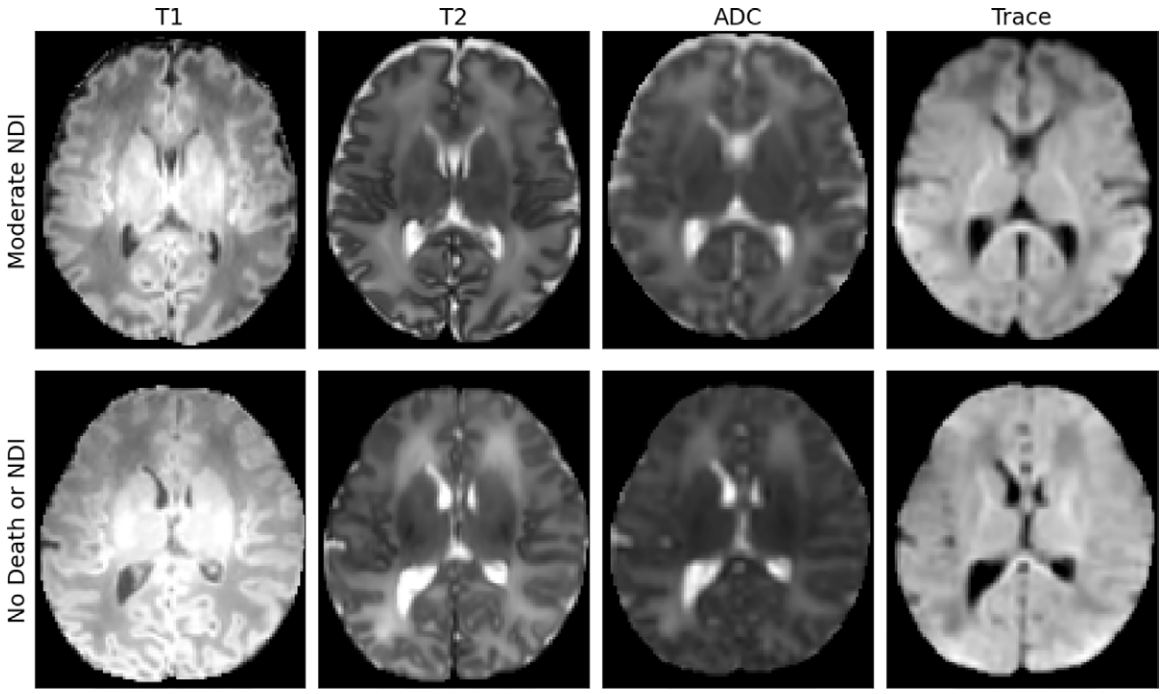

Figure 2: Example MR images for a participant with moderate neurodevelopmental impairment (NDI) (top row) and no death or NDI (bottom row) at 2-year follow-up. Neither set of images reveals any definite focal areas of brain injury, which highlights the difficulty in prognostication based on MRI. ADC = apparent diffusion coefficient.

each of the following secondary outcomes: death or moderate to severe NDI, death or severe NDI, and death alone.

Model Evaluation and Statistical Analysis

Model performance was evaluated on both the in- and out-of-distribution test sets using standard binary classification metrics, including sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUC). Binary classification thresholds were determined using the Youden index for each set of model predictions on the test sets (19). AUC CIs for both test sets were calculated using bootstrapping.

To provide additional context, we compared model performance to the performance of a previously published multireader MRI scoring system that was found to be a predictor of neurodevelopmental outcomes at 18–24 months of age (20). This scoring system used the same imaging sequences to quantify signal abnormality in the following regions: caudate nucleus, globus pallidus and putamen, thalamus, posterior limb of the internal capsule, cerebral white matter, cerebral cortex, cerebellum, and brainstem. Receiver operating characteristic curves were compared using the DeLong method, with $P < .05$ considered statistically significant (21). Uncertainty metrics were calculated using the Monte Carlo dropout technique (22) at the same dropout rate used in training (30%) to examine average variability across predictions, each sampled 30 times.

Model Saliency Evaluation

Image saliency maps for the OPiNE model were generated using the gradients of the final convolutional layer of each single-channel MRI model. Normalized gradients were displayed in a color overlay on the corresponding MR image data to highlight image regions that were potentially important for classification (23).

Results

Characteristics of Study Sample

A total of 500 participants were enrolled in the original HEAL study, and 414 of these participants (mean gestational age, 39 weeks ± 1.4; 232 [56%] males and 182 [44%] females) met the inclusion criteria for our study. Discovery set data consisted of 332 (80%) participants, including 266 in the training set and 66 in the validation set. The testing data consisted of 41 (10%) participants in the in-distribution test set and 41 (10%) participants in the out-of-distribution test set. Clinical data for the study sample are shown in Table 1. Moderate to severe encephalopathy was an inclusion criterion for the original HEAL trial, and of the 414 participants included in our study, 91 (22%) had severe encephalopathy and the remaining 323 (78%) had moderate encephalopathy. A total of 198 (48%) participants experienced the primary outcome of the HEAL trial: death or any NDI. Among these participants, 36 (18%) died at an average age of 11 days (IQR, 6–56 days). Mild NDI was diagnosed in 50 (25%) participants, moderate NDI in 61 (31%), and severe NDI in 51 (26%). Based on apparent diffusion coefficient thresholding at less than 800 × 10^{-6} mm^2/sec, many participants had some degree of acute brain injury, with an average of 12 mL ± 54 of involved brain tissue (17). Additional demographic, clinical, and outcome characteristics of the study sample are shown in Table S1.

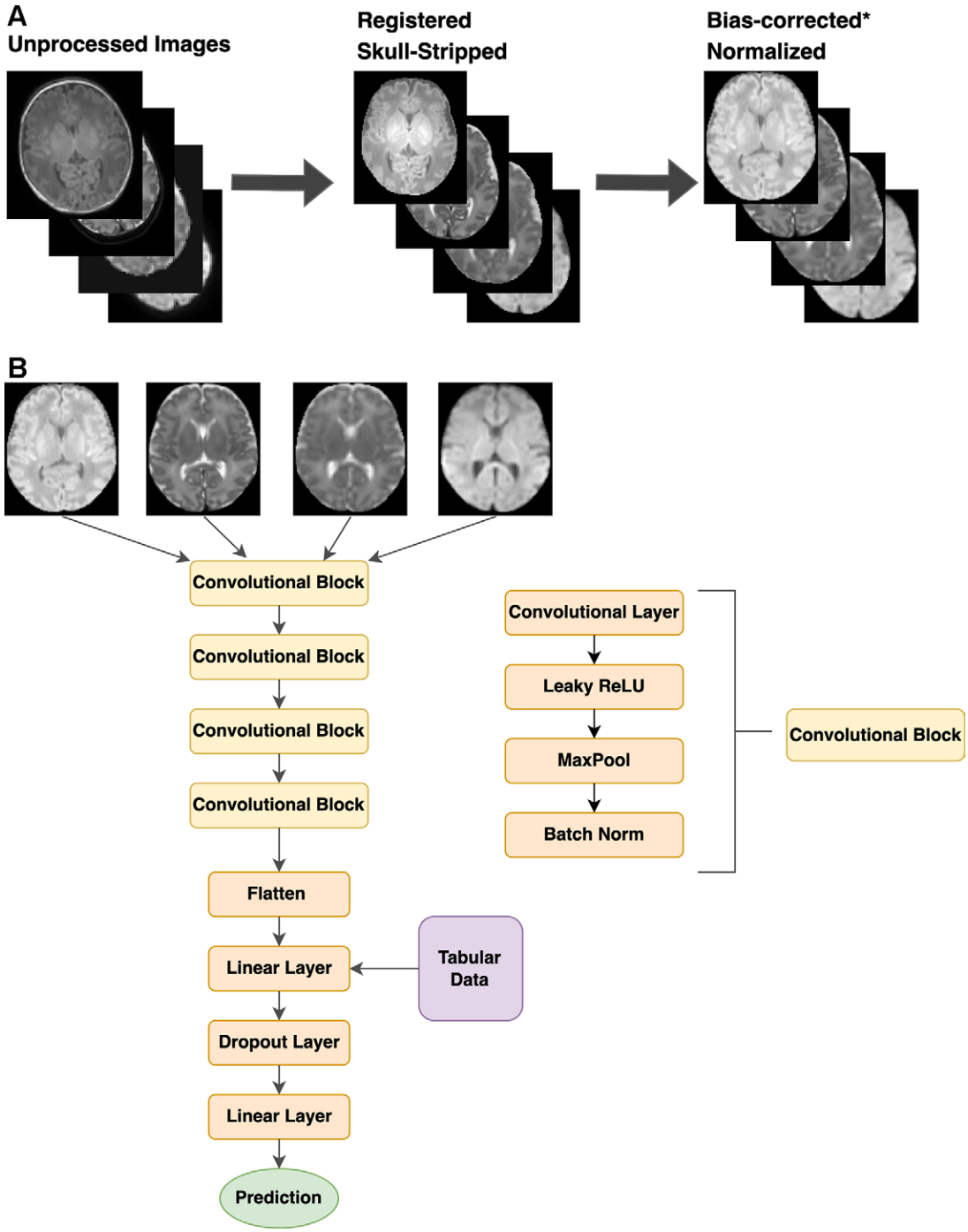

Figure 3: Flowchart of **(A)** imaging preprocessing steps and **(B)** model architecture overview. *Only T1- and T2-weighted images underwent bias correction.

Model Performance

Performance of all models for the primary outcome of death or any NDI is shown in Table 2. Individual MRI sequences achieved AUC values ranging from 0.68 to 0.79 in the in-distribution test set and 0.71 to 0.77 in the out-of-distribution test set. The multisequence OPiNE model achieved an AUC of 0.74 (95% CI: 0.60, 0.86) in the in-distribution test set and 0.77 (95% CI: 0.63, 0.90) in the out-of-distribution test set. When thresholded for binary classification, the OPiNE model yielded an accuracy of 63% in the in-distribution test set and 78% in the out-of-distribution test set. For comparison, logistic regression of tabular data alone yielded an AUC of 0.67 (95% CI: 0.52, 0.82) in the in-distribution test set and 0.58 (95% CI: 0.43, 0.73) in the out-of-distribution test set. The previously published multireader MRI scoring system yielded an AUC of 0.49 (95% CI: 0.32, 0.66) in the in-distribution test set and 0.77 (95% CI: 0.64, 0.88) in the out-of-distribution test set (Fig 4). Additional graphs of receiver operating characteristic curves for all models on both test sets can be found in Figure S1.

Table 1: Clinical Information for Participants in the Current Imaging Analysis

Characteristic	Study Sample (*n* = 414)	Death or NDI (*n* = 198)	No Death or NDI (*n* = 216)
Death or NDI	198 (48)	198 (100)	0
Gestational age (w)	39 ± 1.4	39 ± 1.4	39 ± 1.3
Sex			
Male	232 (56)	122 (62)	110 (51)
Female	182 (44)	76 (38)	106 (49)
Encephalopathy severity			
Moderate	323 (78)	133 (67)	190 (88)
Severe	91 (22)	65 (33)	26 (12)
Outcome			
No death or NDI	216 (52)	0	216 (100)
Mild NDI	50 (12)	50 (25)	0
Moderate NDI	61 (15)	61 (31)	0
Severe NDI	51 (12)	51 (26)	0
Death	36 (9)	36 (18)	0
Brain injury volume (mL)	12 ± 54	25 ± 75	0 ± 3
EPO treatment given	215 (52)	104 (53)	111 (51)

Note.—Continuous variables are presented as means ± SDs. Categorical variables are presented as numbers with percentages in parentheses. Brain injury volume was calculated as the number of voxels with apparent diffusion coefficient values less than 800 × 10^{-6} mm^2/sec. Erythropoietin (EPO) treatment was done as part of the High-dose Erythropoietin for Asphyxia and Encephalopathy study and was not found to impact outcomes. NDI = neurodevelopmental impairment.

Comparing the OPiNE model and logistic regression on tabular data alone, there was no evidence of a difference in the in-distribution test set (P = .35), but the OPiNE model demonstrated superior performance in the out-of-distribution test set (P = .04). Comparing the OPiNE model and the radiologist multireader scoring system, there was no evidence of a difference in performance in either the in-distribution (P = .08) or out-of-distribution (P = .98) test sets. Use of the Monte Carlo technique to estimate uncertainty of the OPiNE model demonstrated an average uncertainty of 0.044 ± 0.059 in the in-distribution test set and 0.033 ± 0.060 in the out-of-distribution test set. Additional model performance statistics for the training and validation sets can be found in Table S2.

Image Saliency Analysis

Saliency maps built from gradient-weighted class activation mapping of each MRI sequence are shown in Figure 5, with examples from participants with and without the primary outcome. Additional saliency maps across the full volume for each MRI sequence and outcome can be found in Figure S2. Gradient hot spots were largely located within the brain and notably involved multiple bilateral cortical regions on T1-weighted and diffusion trace images and subcortical regions, including the thalamus, on T2-weighted and apparent diffusion coefficient images.

Secondary Outcome Performance

Performance of the OPiNE model for secondary outcome prediction is shown in Table 3. When predicting encephalopathy severity, the OPiNE model performed well, with an AUC of 0.75 (95% CI: 0.58, 0.89) in the in-distribution test set and 0.85 (95% CI: 0.72, 0.96) in the out-of-distribution test set. The model also performed well when predicting more severe outcomes, including severe NDI or death, with an AUC of 0.79 (95% CI: 0.50, >0.99) in the in-distribution test set and 0.80 (95% CI: 0.63, 0.95) in the out-of-distribution test set. The strongest performance was observed when predicting death alone, with an AUC of 0.92 (95% CI: 0.85, 0.98) in the in-distribution test set and 0.95 (95% CI: 0.87, >0.99) in the out-of-distribution test set.

Discussion

HIE is the most common cause of neonatal encephalopathy and long-term neurologic disability in neonates. Although MRI plays an important role in the evaluation of neonates with encephalopathy, interpretation can be challenging with variable prognostic value even among experts (14,24). Until recently, there has been relatively little artificial intelligence research in this area, which is likely at least in part related to relative data scarcity. In this study, we leveraged the harmonized MR image data and neurodevelopmental outcome data from the HEAL trial, which enrolled 500 neonates with encephalopathy from 17 sites and followed survivors for 24 months for standardized neurodevelopmental outcome. The harmonized MRI protocol minimizes data heterogeneity without artificially limiting data to a specific patient population, site, or scanner. These data therefore have the potential to yield predictive artificial intelligence models with ideal, yet still generalizable, performance.

The OPiNE model achieved relatively strong performance for predicting the primary outcome of death or any NDI, with

Table 2: Performance of Classification Models and Radiologist Scoring in Predicting Death or NDI in Neonates with Hypoxic-Ischemic Encephalopathy

Data	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
In-distribution test set						
T1 weighted	0.79 (0.67, 0.91)	73 (60, 87) (30/41)	76 (62, 88) (31/41)	72 (58, 86) (30/41)	63 (48, 78) (26/41)	72 (58, 86) (30/41)
T2 weighted	0.70 (0.55, 0.85)	76 (62, 89) (31/41)	44 (29, 59) (18/41)	95 (90, 100) (39/41)	88 (77, 98) (36/41)	95 (90, 100) (39/41)
Apparent diffusion coefficient	0.74 (0.61, 0.87)	66 (51, 80) (27/41)	80 (69, 93) (33/41)	56 (41, 71) (23/41)	54 (39, 69) (22/41)	56 (41, 71) (23/41)
Trace	0.68 (0.54, 0.82)	56 (41, 71) (23/41)	88 (77, 98) (36/41)	37 (21, 51) (15/41)	46 (31, 62) (19/41)	37 (21, 51) (15/41)
Tabular data	0.67 (0.52, 0.82)	71 (57, 85) (29/41)	56 (41, 71) (23/41)	80 (68, 92) (33/41)	63 (50, 79) (26/41)	80 (68, 92) (33/41)
OPiNE model	0.74 (0.60, 0.86)	63 (49, 78) (26/41)	88 (77, 98) (36/41)	49 (33, 63) (20/41)	51 (37, 67) (21/41)	49 (33, 63) (20/41)
Radiologist MRI scoring*	0.49 (0.32, 0.66)	47 (31, 64) (17/36)	83 (73, 96) (30/36)	39 (12, 40) (14/36)	39 (23, 55) (14/36)	25 (12, 40) (9/36)
Out-of-distribution test set						
T1 weighted	0.71 (0.54, 0.86)	80 (68, 93) (33/41)	56 (40, 71) (23/41)	100 (100, 100) (41/41)	100 (100, 100) (41/41)	100 (100, 100) (41/41)
T2 weighted	0.77 (0.64, 0.90)	76 (62, 89) (31/41)	56 (40, 71) (23/41)	90 (83, 100) (37/41)	83 (72, 95) (34/41)	90 (83, 100) (37/41)
Apparent diffusion coefficient	0.73 (0.58, 0.87)	73 (60, 87) (30/41)	66 (52, 81) (27/41)	78 (66, 91) (32/41)	71 (57, 85) (29/41)	78 (66, 91) (32/41)
Trace	0.73 (0.59, 0.88)	76 (62, 89) (31/41)	56 (40, 71) (23/41)	91 (83, 100) (37/41)	83 (72, 95) (34/41)	90 (83, 100) (37/41)
Tabular data	0.58 (0.43, 0.73)	59 (43, 74) (24/41)	66 (52, 81) (27/41)	51 (37, 67) (21/41)	51 (37, 67) (21/41)	51 (37, 67) (21/41)
OPiNE model	0.77 (0.63, 0.90)	78 (65, 91) (32/41)	56 (40, 71) (23/41)	95 (89, 100) (39/41)	90 (82, 100) (37/41)	95 (89, 100) (39/41)
Radiologist MRI scoring*	0.77 (0.64, 0.88)	60 (45, 75) (24/40)	100 (100, 100) (40/40)	28 (13, 41) (11/40)	52 (37, 68) (21/40)	28 (13, 41) (11/40)

Note.—Numbers in parentheses are 95% CIs or numbers of participants. Accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated at a threshold that optimized the difference between true-positive rate and false-positive rate and are presented as percentages. Logistic regression was used to form predictions using tabular data, which included brain injury volume, sex, gestational age, and erythropoietin administration. The Outcome Prediction in Neonates with Encephalopathy (OPiNE) model used T1-weighted, T2-weighted, apparent diffusion coefficient, and trace images combined with tabular data. AUC = area under the receiver operating characteristic curve, NDI = neurodevelopmental impairment.

* The prior study that used reader scoring (Trivedi et al [20]) did not contain five participants in the in-distribution test set and one participant in the out-of-distribution test set.

AUCs above 0.74 for both in- and out-of-distribution test sets. OPiNE model performance was significantly higher compared with logistic regression of tabular data alone in out-of-distribution data, indicating that imaging data contributes important prognostic information. In addition, OPiNE model performance was slightly, but not significantly, higher when compared with a previously published multireader scoring system. Secondary outcome analysis showed similarly strong performance for predicting severe NDI or death and even higher performance for predicting death alone, with AUC values above 0.90. These more severe neurodevelopmental outcomes are potentially more clinically relevant when providing prospective counseling to families. Overall, these results highlight the prognostic utility of neonatal brain MRI and suggest a potential role for artificial intelligence–based outcome prediction to aid in neuroprognostication.

We observed a trend toward superior performance of the OPiNE model in the out-of-distribution test set compared with the in-distribution test set, which is somewhat atypical. There are at least two possible explanations for this observation. First, the out-of-distribution test set was not selected randomly but rather was chosen as a combination of all data from two sites representing approximately 10% of the total dataset and with approximately 50% positivity for the primary outcome. This selection method could have inadvertently resulted in a test set with features that facilitate image-based prediction of outcomes, such as higher-than-average quality imaging or more conspicuous

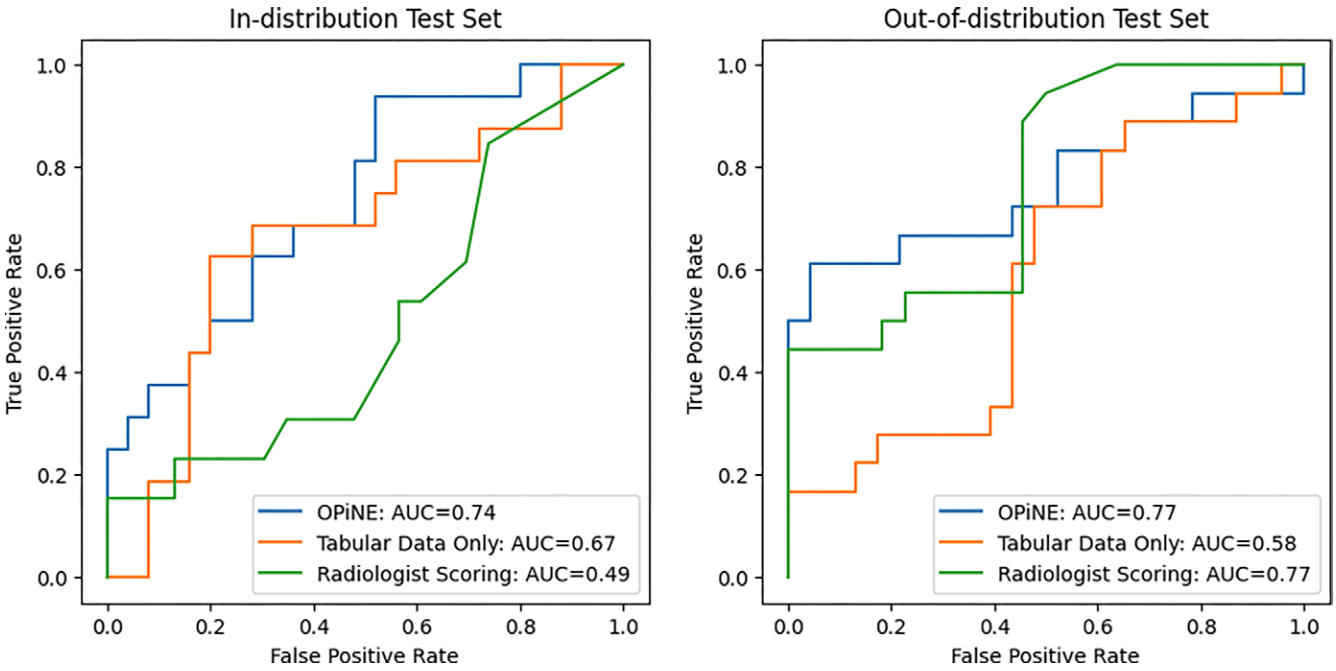

Figure 4: Graph of receiver operating characteristic curves for predictions in the in-distribution and out-of-distribution test sets. The Outcome Prediction in Neonates with Encephalopathy (OPiNE) model used T1-weighted, T2-weighted, apparent diffusion coefficient, trace, and readily available clinical tabular data to perform predictions. The tabular data–only model used logistic regression on the tabular data only. Radiologist scoring used the same imaging sequences as the OPiNE model. All methods were compared using the DeLong method, and there was a difference between the OPiNE and tabular data–only model in the out-of-distribution test set. All other comparisons demonstrated no difference. AUC = area under the receiver operating characteristic curve.

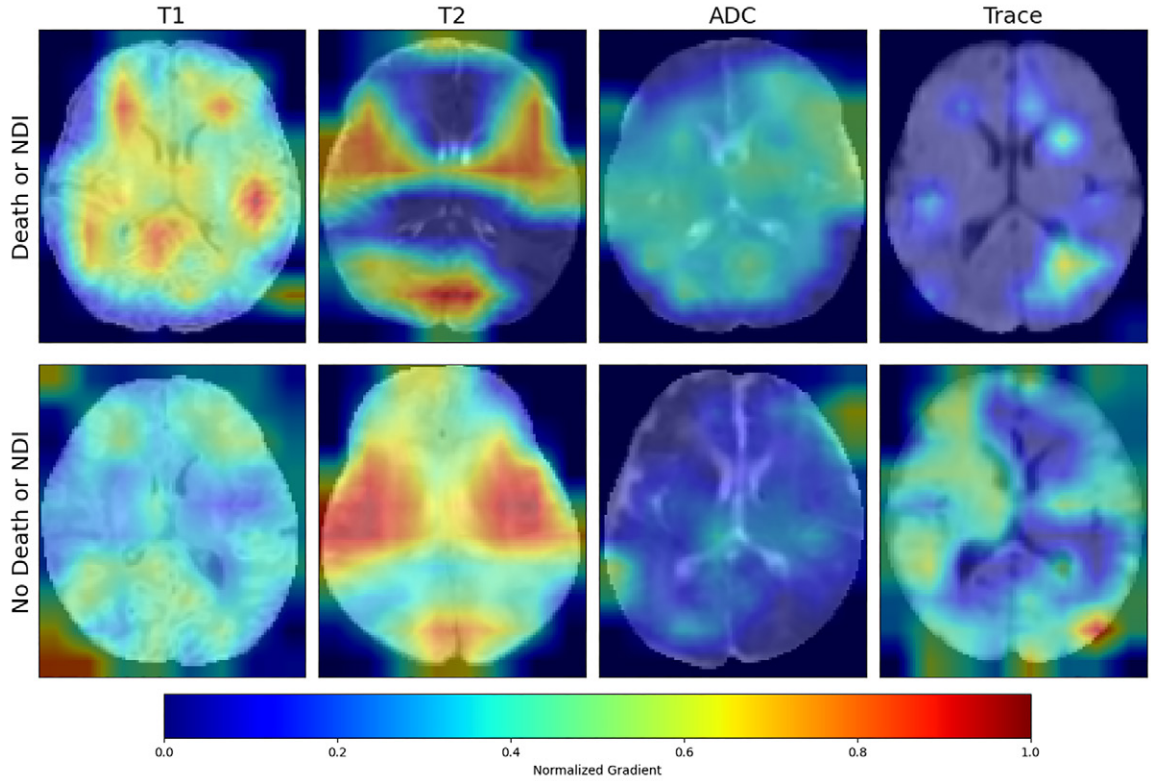

Figure 5: Gradient-weighted class activation mapping overlaid on T1-weighted, T2-weighted, apparent diffusion coefficient (ADC), and trace images at the level of the basal ganglia and thalami for cases of death or neurodevelopmental impairment (NDI, top row) and no death or NDI (bottom row). Gradients of the final convolutional layer were scaled between 0 and 1 and demonstrate salient areas of the image used in classification. Gestational age and sex for each neonate included in the figure, from left to right, are: upper row, 39-week male, 36-week male, 36-week male, and 39-week male; lower row, 40-week male, 39-week female, 41-week male, and 40-week female.

Table 3: OPiNE Performance on Predicting Secondary Outcomes

Target	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
In-distribution test set						
Moderate NDI, severe NDI, or death	0.75 (0.58, 0.89)	73 (60, 87) (30/41)	59 (43, 73) (24/41)	80 (67, 92) (33/41)	54 (39, 69) (22/41)	80 (67, 92) (33/41)
Severe NDI or death	0.79 (0.50, >0.99)	83 (71, 94) (34/41)	61 (45, 75) (25/41)	85 (76, 97) (35/41)	37 (23, 52) (15/41)	85 (76, 97) (35/41)
Death	0.92 (0.85, 0.98)	90 (81, 99) (37/41)	49 (35, 65) (20/41)	93 (84, 100) (38/41)	24 (12, 38) (10/41)	93 (84, 100) (38/41)
Out-of-distribution test set						
Moderate NDI, severe NDI, or death	0.85 (0.72, 0.96)	85 (75, 96) (35/41)	61 (47, 76) (25/41)	98 (91, 100) (40/41)	88 (79, 99) (36/41)	98 (91, 100) (40/41)
Severe NDI or death	0.80 (0.63, 0.95)	76 (62, 89) (31/41)	66 (52, 81) (27/41)	78 (65, 91) (32/41)	46 (31, 61) (19/41)	78 (65, 91) (32/41)
Death	0.95 (0.87, >0.99)	90 (81, 99) (37/41)	66 (52, 81) (27/41)	93 (84, 100) (38/41)	39 (25, 55) (16/41)	93 (84, 100) (38/41)

Note.—The final linear layer of the model was fine-tuned using training and validation set data. Numbers in parentheses are 95% CIs or numbers of participants. Accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are presented as percentages. AUC = area under the receiver operating characteristic curve, NDI = neurodevelopmental impairment, OpiNE = Outcome Prediction in Neonates with Encephalopathy.

imaging findings in participants with death or any NDI. Second, it is possible that the OPiNE model has a relatively high variability, either due to the difficult nature of the task or the design of the model itself. This potential variability could explain the trend toward higher performance in the out-of-distribution test set, which would typically be similar or lower compared with an in-distribution test set for a low variability model. To address this, the OPiNE model should be evaluated on larger and more diverse cohorts, and we have made the model publicly available to facilitate this process. Overall, the out-of-distribution test set results are encouraging for generalizability of the OPiNE model, but further validation using site-specific data should be considered a prerequisite before any potential clinical implementation.

We performed a gradient-weighted class activation mapping analysis to provide basic insight into the OPiNE model, and although gradient analysis has known limitations, it provides a general sense of salient regional image features. Many gradient hot spots were identified within brain regions that are commonly affected in hypoxic ischemic injury. For example, the basal ganglia and thalamic pattern of hypoxic-ischemic injury primarily affects the perirolandic cortex, corticospinal tract, and deep gray nuclei, whereas the peripheral and watershed pattern tends to affect the cortex and subcortical white matter (25–27). In our saliency maps, gradient activation was commonly observed in the white matter on T2-weighted images, which became increasingly T2 hyperintense with increasing injury severity. Additionally, gradient activation was commonly observed centrally within the basal ganglia, especially the thalami, on diffusion-weighted images in those who died or had NDI, which is consistent with prior observations that thalamic injury is associated with poor outcomes (17). By contrast, gradient activation was more common in the peripheral subcortical white matter on diffusion-weighted trace images among those with no death or NDI, supporting the hypothesis that these watershed pattern injuries less commonly lead to substantial NDI by 24-month follow-up (28,29).

It is important to note that the HEAL trial enrolled participants who were born at 36 weeks of gestation or older and who had moderate or severe encephalopathy by Sarnat criteria at between 1 and 6 hours of age. This is the population of infants who currently qualify for therapeutic hypothermia. As the imaging and outcomes in the trial are derived from this population, the results of our model are not generalizable beyond this population. Thus, infants with neonatal encephalopathy outside these criteria, such as those who have mild encephalopathy by Sarnat criteria or those who develop encephalopathy beyond 6 hours of life, fall outside the realm of prediction from our model. Several previous studies have examined the effectiveness of MRI outcome prognostication for neonates with HIE using radiologist interpretation and scoring systems. One study demonstrated that radiologists perform similarly to our models (AUC of 0.77), though with a smaller study sample of 128 participants (20.3% with poor outcome) (30). Another related study with a more comparable sample size of 486 participants achieved a high AUC for predicting death or NDI (0.85) but included a combination of MRI, electroencephalography, and several clinical and treatment variables (31). Several other studies have demonstrated notable neuroprognostic performance but with additional inclusion criteria, such as requiring certain MRI findings (32) or predicting death alone (33,34). Compared with prior work, our study is notable for the relatively large multi-institutional sample, harmonized imaging protocol that included real-world variation in vendors and platforms, and well-defined 24-month neurodevelopmental outcomes. In addition, the OPiNE model uses only T1-, T2-, and diffusion-weighted images, which are commonly acquired in clinical practice.

This study had several limitations. First, the sample size of 414 participants, while large in the context of neonatal HIE,

is relatively small compared with other medical imaging deep learning cohorts. We used data augmentation and multiple evaluation methods with an in-distribution and out-of-distribution test set to provide robust results despite the limited dataset. Second, only a relatively small set of clinical variables was included in the model. Although this may limit the performance of the model, removing additional clinical variables simplifies model deployment as all data needed for model use are included in most picture archiving and communication systems. Third, the data used in this study originated from a clinical trial in which half of all participants received erythropoietin as an adjunct to therapeutic hypothermia in the setting of neonatal encephalopathy. The HEAL trial ultimately concluded no therapeutic benefit compared with placebo, and thus we believe that erythropoietin administration is not a confounder in this study. Nonetheless, we have adjusted for erythropoietin administration as part of the tabular data of this study. Fourth, although the fine-tuned OPiNE model performed strongly at predicting death, these results should be viewed in the context of a small sample size with death as an outcome. Finally, it is important to note that many model statistics, including sensitivity, specificity, negative predictive value, and positive predictive value, can be altered by changing the model's threshold. We used the Youden index to determine an optimal threshold in our study, but clinical practice may require higher sensitivity at the cost of specificity. An appropriate model threshold may also require a larger, more robust dataset to allow for generalizability. However, regardless of threshold choice, image-based predictive models are inherently uncertain, and OPiNE should be considered a prognostic tool rather than a method to determine patient outcomes.

In conclusion, our results indicate that artificial intelligence may be able to use neonatal brain MRI to effectively predict 2-year neurodevelopmental outcomes. Further work with a larger dataset or combining additional readily available clinical information may further confirm and expand on our study.

Acknowledgments: We thank the members of the HEAL Consortium, listed in the supplementary Appendix of the HEAL study *(https://doi.org/10.1056/nejmoa2119660)* (9), for their role in gathering the data used in this study.

Author contributions: Guarantors of integrity of entire study, **C.O.L., Y.L.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **C.O.L., E.C., F.T., A.L., Y.L.**; clinical studies, **C.O.L., E.C., A.L., S.J., J.L.W., Y.W.W., Y.L.**; experimental studies, **C.O.L., J.V.C., R.C.M., A.R.**; statistical analysis, **C.O.L., Y.L.**; and manuscript editing, **C.O.L., E.C., J.V.C., F.T., G.C., A.L., J.F., S.J., A.M., R.C.M., A.R., Y.W.W., Y.L.**

Data sharing: Data analyzed during the study were provided by a third party. Requests for data should be directed to the provider indicated in the Acknowledgments.

Disclosures of conflicts of interest: **C.O.L.** No relevant relationships. **E.C.** No relevant relationships. **J.V.C.** No relevant relationships. **F.T.** No relevant relationships. **G.C.** No relevant relationships. **A.L.** No relevant relationships. **J.F.** No relevant relationships. **S.J.** Support for the present article from the National Institutes of Health (NIH) National Institute of Neurological Disorders and Stroke (NINDS) grants 1U01NS092764 and U01NS092553, paid to author's institution; grants or contracts from the NINDS (1R13NS127525-01, 2023-2024 and R01HD101422-01A1, 2021-2026), the National Institute of Child Health and Human Development (NICHD) (P50 HD103524, 2020-2025 and 1R01HD107003-01, 2022-2027), and COOL Prime (2022), all paid to author's institution; royalties from Elsevier for editing *Avery's Diseases of the Newborn*, 10th edition, paid to author directly; support from the above-mentioned grants for attending meetings and/or travel, paid to author's institution; participation on a Data Safety monitoring board or advisory board for ALBINO, COOL Prime, and 1K23HL150300-01A1 (Enteral iron supplementation and intestinal health in preterm infants), no payment; director of the Institute on Human Development and Disability, paid position, paid to author directly. **A.M.** NIH grant RO1: HEAL trial. **R.C.M.** Grant 5U01NS092764 High-dose Erythropoietin for Asphyxia and Encephalopathy (HEAL), payments made to Washington University; support from Siemens Healthcare and Philips Healthcare for travel and meals, to learn about MRI and CT scanners (author helps make purchasing decisions for health care system, unrelated to the research in this article); stock options from Turing Medical for medical advisory board participation (Turing makes software and hardware for MRI scanners, their products were not used in this article). **J.L.W.** Support for the present article provided by the NIH (U01NS092764 and U01NS092553), author received no additional funds toward this work. **A.R.** Grant support from GE HealthCare, the Hydrocephalus Association, UCSF Helen Diller Comprehensive Cancer Center, NCCN, and the Society for Pediatric Radiology; consulting fees from Arterys; stock options for MRImatch. **Y.W.W.** NIH grant U01NS092764; participation on the NICHD Neonatal Research Network Data Safety Monitoring Committee. **Y.L.** NIH grant U01 NS092764-01; leadership or fiduciary role on the American Society of Pediatric Radiology Board of Directors.

References

1. Lawn JE, Kerber K, Enweronu-Laryea C, Cousens S. 3.6 million neonatal deaths--what is progressing and what is not? Semin Perinatol 2010;34(6):371–386.
2. Heinz ER, Provenzale JM. Imaging findings in neonatal hypoxia: a practical review. AJR Am J Roentgenol 2009;192(1):41–47.
3. Miller SP, Ramaswamy V, Michelson D, et al. Patterns of brain injury in term neonatal encephalopathy. J Pediatr 2005;146(4):453–460.
4. Dixon BJ, Reis C, Ho WM, Tang J, Zhang JH. Neuroprotective Strategies after Neonatal Hypoxic Ischemic Encephalopathy. Int J Mol Sci 2015;16(9):22368–22401.
5. Goergen SK, Ang H, Wong F, et al. Early MRI in term infants with perinatal hypoxic-ischaemic brain injury: interobserver agreement and MRI predictors of outcome at 2 years. Clin Radiol 2014;69(1):72–81.
6. Baker S, Kandasamy Y. Machine learning for understanding and predicting neurodevelopmental outcomes in premature infants: a systematic review. Pediatr Res 2023;93(2):293–299.
7. Weiss RJ, Bates SV, Song Y, et al. Mining multi-site clinical data to develop machine learning MRI biomarkers: application to neonatal hypoxic ischemic encephalopathy. J Transl Med 2019;17(1):385.
8. Juul SE, Comstock BA, Heagerty PJ, et al. High-Dose Erythropoietin for Asphyxia and Encephalopathy (HEAL): A Randomized Controlled Trial - Background, Aims, and Study Protocol. Neonatology 2018;113(4):331–338.
9. Wu YW, Comstock BA, Gonzalez FF, et al. Trial of Erythropoietin for Hypoxic–Ischemic Encephalopathy in Newborns. N Engl J Med 2022;387(2):148–159.
10. Juul SE, Voldal E, Comstock BA, et al. Association of High-Dose Erythropoietin With Circulating Biomarkers and Neurodevelopmental Outcomes Among Neonates With Hypoxic Ischemic Encephalopathy: A Secondary Analysis of the HEAL Randomized Clinical Trial. JAMA Netw Open 2023;6(7):e2322131.
11. Juul SE, McPherson RJ, Bammler TK, Wilkerson J, Beyer RP, Farin FM. Recombinant erythropoietin is neuroprotective in a novel mouse oxidative injury model. Dev Neurosci 2008;30(4):231–242.
12. Kuban KCK, Allred EN, O'Shea M, et al. An algorithm for identifying and classifying cerebral palsy in young children. J Pediatr 2008;153(4):466–472.
13. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997;39(4):214–223.
14. Wu YW, Monsell SE, Glass HC, et al. How well does neonatal neuroimaging correlate with neurodevelopmental outcomes in infants with hypoxic-ischemic encephalopathy? Pediatr Res 2023;94(3):1018–1025.
15. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 2016;264:47–56.
16. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.
17. Calabrese E, Wu Y, Scheffler AW, et al. Correlating Quantitative MRI-based Apparent Diffusion Coefficient Metrics with 24-month Neurodevelopmental Outcomes in Neonates from the HEAL Trial. Radiology 2023;308(3):e223262.

18. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015; 1026–1034.
19. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–35.
20. Trivedi SB, Vesoulis ZA, Rao R, et al. A validated clinical MRI injury scoring system in neonatal hypoxic-ischemic encephalopathy. Pediatr Radiol 2017;47(11):1491–1499.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
22. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ArXiv 1506.02142 [preprint] https://arxiv.org/abs/1506.02142. Posted June 6, 2015. Accessed March 20, 2024.
23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017; 618–626.
24. Langeslag JF, Berendse K, Daams JG, et al. Clinical Prediction Models and Predictors for Death or Adverse Neurodevelopmental Outcome in Term Newborns with Hypoxic-Ischemic Encephalopathy: A Systematic Review of the Literature. Neonatology 2023;120(6):776–788.
25. Barkovich A, Raybaud C. Brain and Spine Injuries in Infancy and Childhood. In: Pediatric Neuroimaging. 6th ed. Wolters Kluwer, 2019.
26. Barkovich AJ. MR and CT evaluation of profound neonatal and infantile asphyxia. AJNR Am J Neuroradiol 1992;13(3):959–972; discussion 973–975.
27. Okereafor A, Allsop J, Counsell SJ, et al. Patterns of brain injury in neonates exposed to perinatal sentinel events. Pediatrics 2008;121(5):906–914.
28. Harteman JC, Groenendaal F, Toet MC, et al. Diffusion-weighted imaging changes in cerebral watershed distribution following neonatal encephalopathy are not invariably associated with an adverse outcome. Dev Med Child Neurol 2013;55(7):642–653.
29. Lee BL, Glass HC. Cognitive outcomes in late childhood and adolescence of neonatal hypoxic-ischemic encephalopathy. Clin Exp Pediatr 2021;64(12):608–618.
30. Laptook AR, Shankaran S, Barnes P, et al. Limitations of Conventional Magnetic Resonance Imaging as a Predictor of Death or Disability Following Neonatal Hypoxic-Ischemic Encephalopathy in the Late Hypothermia Trial. J Pediatr 2021;230:106–111.e6.
31. Peeples ES, Rao R, Dizon MLV, et al. Predictive Models of Neurodevelopmental Outcomes After Neonatal Hypoxic-Ischemic Encephalopathy. Pediatrics 2021;147(2):e2020022962.
32. Martinez-Biarge M, Diez-Sebastian J, Kapellou O, et al. Predicting motor outcome and death in term hypoxic-ischemic encephalopathy. Neurology 2011;76(24):2055–2061.
33. Basiri B, Sabzehei M, Sabahi M. Predictive factors of death in neonates with hypoxic-ischemic encephalopathy receiving selective head cooling. Clin Exp Pediatr 2021;64(4):180–187.
34. Dathe A-K, Stein A, Bruns N, et al. Early Prediction of Mortality after Birth Asphyxia with the nSOFA. J Clin Med 2023;12(13):4322.