Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Development and Evaluation of Deep Learning Models for Automated Estimation of Myelin Maturation Using Pediatric Brain MRI Scans

Tugba Akinci D'Antonoli, MD • Ramona-Alexandra Todea, MD • Nora Leu, MD • Alexandre N. Datta, MD • Bram Stieltjes, MD, PhD • Friederike Pruefer, MD • Jakob Wasserthal, PhD

From the Department of Pediatric Radiology (T.A.D., N.L., F.P.) and Department of Pediatric Neurology and Developmental Medicine (A.N.D.), University Children's Hospital Basel, Spitalstrasse 33, 4056 Basel, Switzerland; Institute of Radiology and Nuclear Medicine, Cantonal Hospital Basel, Basel, Switzerland (T.A.D.); and Department of Neuroradiology, Clinic of Radiology and Nuclear Medicine (R.A.T.) and Department of Research and Analysis, Clinic of Radiology and Nuclear Medicine (B.S., J.W.), University Hospital Basel, Basel, Switzerland. Received December 19, 2022; revision requested February 15, 2023; revision received June 20; accepted July 7. **Address correspondence to** T.A.D. (email: *tugba.akincidantonoli@unibas.ch*).

Authors declared no funding for this work.

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

Radiology: Artificial Intelligence 2023; 5(5):e220292 • https://doi.org/10.1148/ryai.220292 • Content codes: AI MR NR PD

Purpose: To predict the corresponding age of myelin maturation from brain MRI scans in infants and young children by using a deep learning algorithm and to build upon previously published models.

Materials and Methods: Brain MRI scans acquired between January 1, 2011, and March 17, 2021, in our institution in patients aged 0–3 years were retrospectively retrieved from the archive. An ensemble of two-dimensional (2D) and three-dimensional (3D) convolutional neural network models was trained and internally validated in 710 patients to predict myelin maturation age on the basis of radiologist-generated labels. The model ensemble was tested on an internal dataset of 123 patients and two external datasets of 226 (0–25 months of age) and 383 (0–2 months of age) healthy children and infants, respectively. Mean absolute error (MAE) and Pearson correlation coefficients were used to assess model performance.

Results: The 2D, 3D, and 2D-plus-3D ensemble models showed MAE values of 1.43, 2.55, and 1.77 months, respectively, on the internal test set, values of 2.26, 2.27, and 1.22 months on the first external test set, and values of 0.44, 0.27, and 0.31 months on the second external test set. The ensemble model outperformed the previous state-of-the-art model on the same external test set (MAE = 1.22 vs 2.09 months).

Conclusion: The proposed deep learning model accurately predicted myelin maturation age using pediatric brain MRI scans and may help reduce the time needed to complete this task, as well as interobserver variability in radiologist predictions.

Supplemental material is available for this article.

© RSNA, 2023

Myelin maturation is a dynamic process that begins before birth and continues during the first 3 years of life (1,2). At birth, myelination is present in the white matter of the cerebellum and brainstem, posterior limb of the internal capsule, ventrolateral thalami, and subcortical white matter of primary motor and sensory cortices (Fig 1). After birth, myelination progresses from central to peripheral, caudal to rostral, dorsal to ventral, and sensory to motor (3). Therefore, determining normal myelin development in the ever-changing brains of infants and young children often requires years of expertise and is considered challenging even by experts in the field of pediatric neuroradiology. Furthermore, despite years of experience, this assessment process is inherently complex and subject to variation both within and between observers.

In the daily workflow of pediatric radiology, radiologists assess myelin maturation by reviewing age-appropriate milestones of normal myelinization (4) and comparing images in patients with images in normal developmental atlases (5). Automating and objectifying this step could support and hasten the clinical decision-making process. To date, there have been few attempts to predict chronologic age using brain MRI, but such studies were conducted on small sample sizes composed of either adults or individuals through late childhood (6–11).

A deep learning algorithm may be useful in predicting normal age-specific myelin development on the basis of brain MRI. Such a pipeline could help overcome intra- and interobserver variability while aiding the clinical decision-making process. In further steps, the same method could be used to identify patients with developmental delays and predict outcome. In this study, we aim to develop a deep learning model to estimate corresponding myelin maturation age on pediatric brain MRI scans. In addition, we aim to externally test and build upon a previously published model (11) while providing a large and diverse open-access dataset.

Materials and Methods

This retrospective study was approved by the Ethics Committee Northwest and Central Switzerland (EKNZ BASEC identification no. 2021–01682) with waiver of informed consent.

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

CNN = convolutional neural network, dHCP = Developing Human Connectome Project, MAE = mean absolute error, NIH = National Institutes of Health, PACS = picture archiving and communication system, 3D = three-dimensional, 2D = two-dimensional

Summary

An ensemble of two-dimensional and three-dimensional convolutional neural network models accurately predicted radiologist-determined myelin maturation age on T1- and T2-weighted brain MRI scans in infants and young children.

Key Points

- The proposed deep learning approach used an ensemble of two-dimensional and three-dimensional convolutional neural networks to predict myelin maturation from MRI scans.
- The proposed ensemble approach demonstrated good performance, with a small mean absolute error in the internal test set (1.77 [95% CI: 1.37, 2.23] and two external test sets (1.22 [95% CI: 1.08, 1.37] and 0.31 [95% CI: 0.29, 0.33]).
- The model prediction had a strong agreement with the ground truth in the internal test set (concordance coefficient = 0.96 [95% CI: 0.94, 0.98]) and first external set (0.96 [95% CI: 0.95, 0.97]) and satisfactory agreement in the second external test set (concordance coefficient = 0.52 [95% CI: 0.46, 0.57]).

Keywords

Pediatrics, MR Imaging, CNS, Brain/Brain Stem, Convolutional Neural Network (CNN), Artificial Intelligence, Pediatric Imaging, Myelin Maturation, Brain MRI, Neuroradiology

Data Mining

All data were retrospectively retrieved from the picture archiving and communication system (PACS) using a search tool, PACS/RISCrawler (version 1.0; *https://zenodo.org/record/3579239*) (12), which allows for both a free-text search and a more advanced search in the PACS. In this search tool, a query adapted from the study's inclusion criteria was used, as follows: *(a)* brain MRI examinations, *(b)* patients aged from 0 to 3 years, and *(c)* examinations between January 1, 2011, and March 17, 2021. Exclusion criteria were then applied, as follows: *(a)* images transferred from external institutions, *(b)* patients older than 36 months of age, and *(c)* follow-up examinations for the same patient. Patients were also excluded manually if they had major pathologic conditions covering almost the entire brain area or if their images were severely distorted because of motion artifacts (Fig S1). Images with minor pathologic features or motion artifacts were left in the dataset to maintain a real-world scenario for model training and testing. The patient selection workflow is shown in Figure 2. The entire dataset is openly available at *https://zenodo.org/record/8055666*.

Image Acquisition

All studies were performed with a Siemens Magnetom Skyra 3-T MRI scanner (Siemens Healthineers) and a 20-channel head coil. We selected axial T1-weighted and T2-weighted native images from each study. Slightly different protocols have been used over the 10 years, and patients were scanned either with or without sedation. More details about image acquisition can be found in Appendix S1.

Image Preprocessing

Selected images were retrieved, de-identified, and converted to Neuroimaging Informatics Technology Initiative format. The T1-weighted images were rigidly registered to the T2-weighted images. A skull stripping algorithm, SynthStrip, was used to extract only the brain area (13) and to avoid re-identification of the patient (14). All images were registered to an 18-month-old healthy reference child by using affine registration. In this way, the model was prevented from learning the age-dependent head size. More details about image preprocessing can be found in Appendix S1.

Data Labeling

All evaluations are carried out on postprocessed T1-weighted and T2-weighted images on the Nora imaging platform *(https://www.nora-imaging.com/)*. The raters (T.A.D., R.A.T.) labeled the myelin development age of the patients who corresponded to the images by evaluating entire brain images and both T1- and T2-weighted images and referred to the same charts (4) and atlases (5) when necessary. Patients born at 37 weeks gestational age or later were considered as term infants and before 37 weeks gestational age as preterm infants. If a patient was preterm, the corrected gestational age was calculated manually by subtracting the number of weeks that the patient was born before 37 weeks from the age (in weeks) at the time of the imaging. See Appendix S1 for further details about data labeling.

Model Training and Cross-Validation

A convolutional neural network (CNN) ensemble was trained by using radiologist-defined myelin age labels. The ensemble was made up of two models, as follows: First, we reimplemented the three-dimensional (3D) CNN from Chen et al (11). To improve its performance, we combined this model with a complementary second model, a two-dimensional (2D) EfficientNet-b0 (15) pretrained with "noisy student self-learning" (16). More details about model training and cross-validation can be found in Appendix S1.

The training of each 3D or 2D model took approximately 1 hour 30 minutes (total training time, 45 hours for 30 models), and inference with the final ensemble for one sample took 1.19 seconds. All models were developed using PyTorch Lightning (17) and trained on one NVIDIA GeForce RTX 3090 GPU. Figure 3 shows the entire pipeline. To ensure open science practices and reproducibility, our code is openly available at *https://github.com/wasserth/MyelinAge*.

External Testing

We used two public datasets for external testing, the National Institutes of Health (NIH) pediatric brain MRI database and the Developing Human Connectome Project (dHCP) database. The same image preprocessing method used for the internal datasets was applied to the external datasets. More details about external datasets can be found in Appendix S1.

Statistical Analysis

G*Power version 3.1 (The G*Power Team) was used to determine the sample size (18). A recent study reported a median

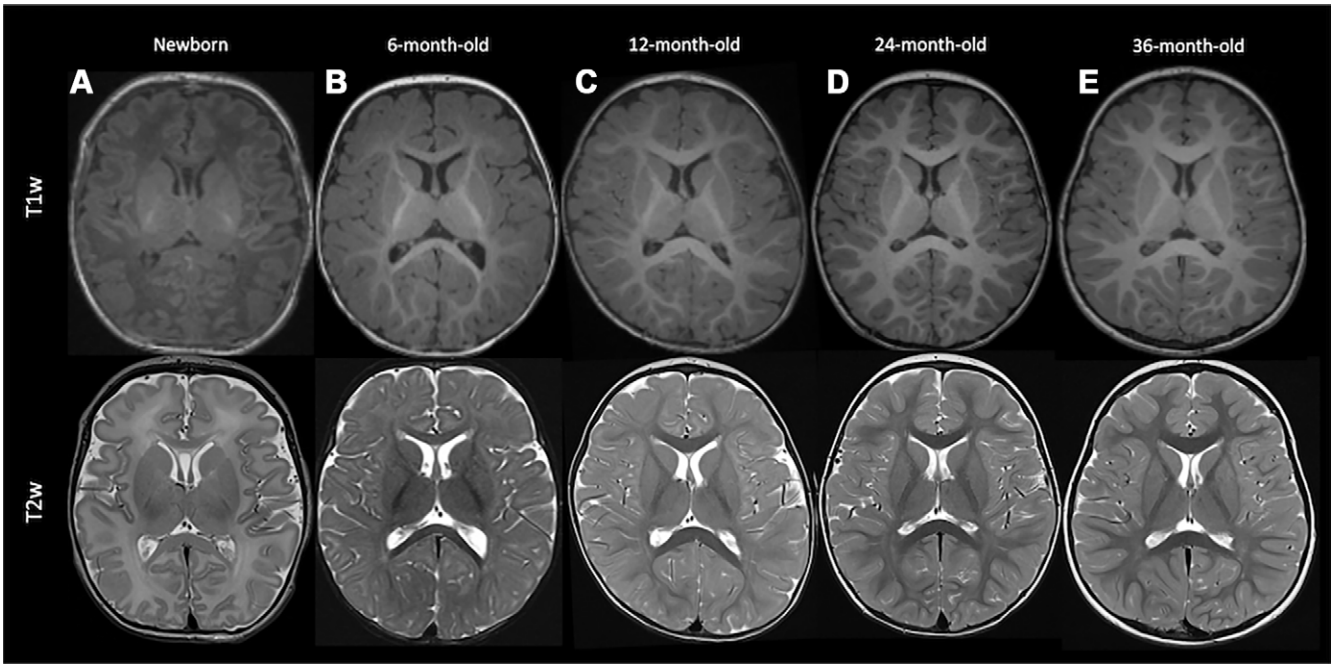


Figure 1: Axial T1- and T2-weighted (T1w and T2w, respectively) MRI scans in infants and young children with normal myelin maturation at the level of the corpus callosum and lateral ventricles. Note that the images are resized to keep the focus only on the progression of myelin maturation at different months. **(A)** Images in a term newborn with meconium aspiration syndrome who underwent imaging to rule out ischemia, infarction, or hemorrhage. **(B)** Images in a 6-month-old infant with a history of asphyxia and sinus vein thrombosis at birth but with normal development. **(C)** Images in a 12-month-old infant with an ocular hemangioma who underwent imaging to rule out an intracranial pathologic condition. **(D)** Images in a 24-month-old child with headache for 2 months. **(E)** Images in a 36-month-old child with a history of focal seizures; imaging was performed to rule out structural changes.

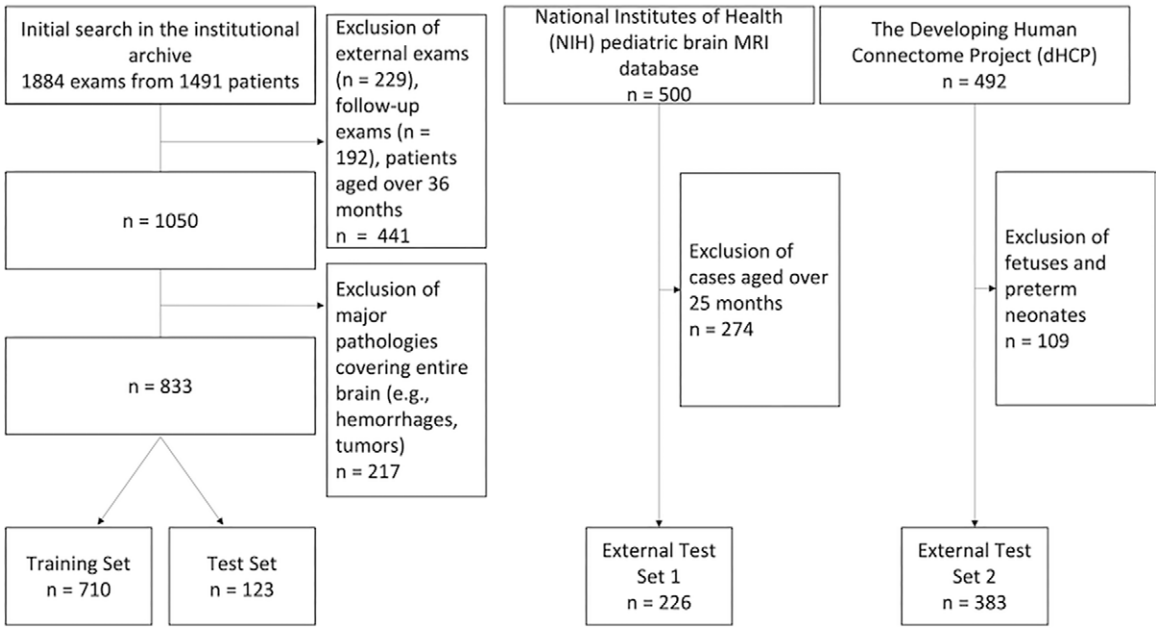


Figure 2: Study flowchart.

difference of 1.9 months ± 0.4 (SD) between the estimated age and the corrected chronologic age (11). Therefore, a sample size of 833 was sufficient to reach a minimum 80% power and 5% error rate according to a two-sided *t* test.

Model performance was evaluated in terms of the absolute error in months (mean absolute error, MAE). Moreover, Lin concordance correlation coefficient (CCC), coefficient of determination (R^2), and Pearson correlation coefficient *(r)* were also calculated. The variability between the ground truth and the model predictions of corresponding myelin developmental age in different datasets was visualized by using Bland-Altman plots. For each metric, we provided 95% CIs, which were calculated by bootstrapping (nonparametric percentile bootstrapping with 10 000 iterations). Python (version 3.7) and the sklearn and

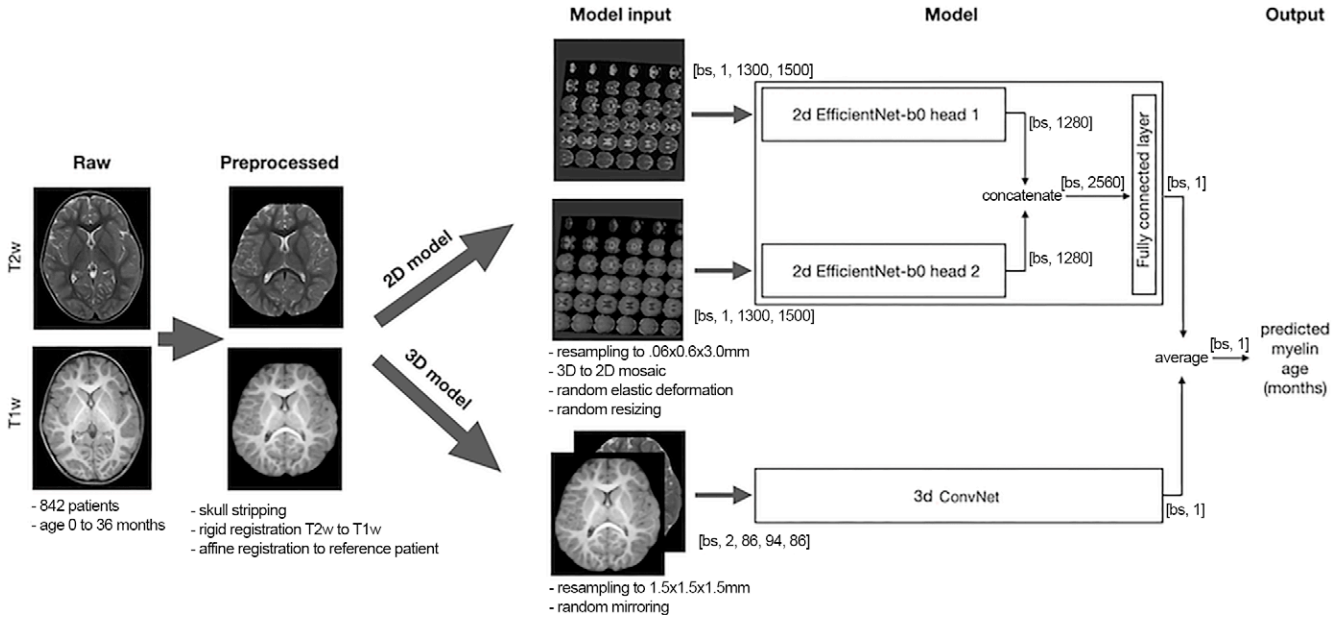


Figure 3: Pipeline of the proposed model. The numbers in square brackets show the dimensions of the input and output at the respective points of the pipeline. bs = batch size, T1w = T1-weighted, T2w = T2-weighted, 3D = three-dimensional, 2D = two-dimensional.

Table 1: Demographics of the Study Sample

Parameter	Training Set (n = 710)	Test Set (n = 123)	dHCP (n = 383)	NIH (n = 226)	P Value
Diagnosis					<.001*
Healthy	540 (76.1)	26 (21.1)	383 (100)	226 (100)	
Premature	22 (3.1)	5 (4.1)	0	0	
Congenital malformation	15 (2.1)	7 (5.7)	0	0	
Hemorrhage	15 (2.1)	9 (7.3)	0	0	
Tumor	4 (0.6)	9 (7.3)	0	0	
Ischemia or infarct	43 (6.1)	41 (33.3)	0	0	
Infection	8 (1.1)	12 (9.8)	0	0	
Developmental condition	45 (6.3)	8 (6.5)	0	0	
Miscellaneous	18 (2.5)	6 (4.9)	0	0	
Age (mo)†	14.1 ± 11.7 (–1, 36)	11.1 ± 11.7 (–1, 36)	0.7 ± 0.4 (0.02, 1.63)	10.4 ± 6 (0.26, 23.84)	<.001‡

Note.—Pathologic conditions are listed by major subgroups. Data in parentheses are percentages unless otherwise indicated. dHCP = Developing Human Connectome Project, NIH = National Institutes of Health.
* χ^2 test.
† Data are means ± SDs, with minimums and maximums in parentheses.
‡ Student t test.

scipy packages were used to conduct all analyses. We used t test for continuous variables (ie, age) and χ^2 test for categorical variables (ie, diagnosis). The α level was set at .05.

Results

Dataset Characteristics

After applying all exclusion criteria, the dataset included 833 infants and children (mean age, 13.7 months ± 11.8). In total, 275 patients had various pathologic features in their examinations (Table 1). Then we did a random data split (85:15). Later, the pediatric radiology fellow (T.A.D.) moved a small number of cases with different pathologic conditions from the training set to the test set, and some healthy patients from the test set to the training set, to ensure that the test set was as representative as possible of the population and to avoid overly optimistic results in the test set (19). As a result, we allocated 85% (710) of the cases for training and cross-validation (fivefold cross-validation for hyperparameter optimization) and 15% (123) of the cases for internal testing. The allocation of every case can be found in the data repository *(https://zenodo.org/record/8055666)*. Figure 4 shows the age distribution of the training set and internal and external test sets.

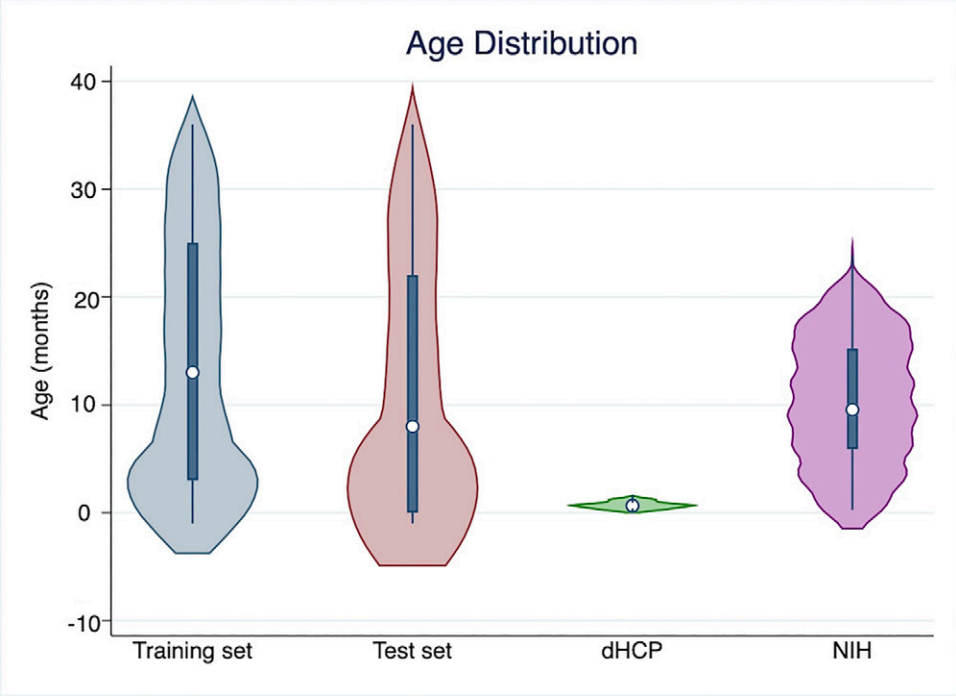


Figure 4: Violin plots show the age distribution of the training set, internal test set, and external test sets. dHCP = Developing Human Connectome Project, NIH = National Institutes of Health.

Model Performance

The 2D model had MAE values of 1.53 months (95% CI: 1.41, 1.66) on the cross-validation set, 1.43 months (95% CI: 1.08, 1.84) on the internal test set, 2.26 months (95% CI: 2.07, 2.48) on the external NIH dataset, and 0.44 months (95% CI: 0.42, 0.47) on the external dHCP dataset.

The 3D model showed MAE values of 2.06 months (95% CI: 1.9, 2.24) on the cross-validation set, 2.55 months (95% CI: 2.01, 3.17) on the internal test set, 2.27 months (95% CI: 1.98, 2.58) on the external NIH dataset, and 0.27 months (95% CI: 0.25, 0.29) on the external dHCP dataset.

The ensemble model (3D + 2D) showed MAE values of 1.63 months (95% CI: 1.5, 1.77) on the cross-validation set, 1.77 months (95% CI: 1.37, 2.23) on the internal test set, 1.22 months (95% CI: 1.08, 1.37) on the external NIH dataset, and 0.31 months (95% CI: 0.29, 0.33) on the external dHCP dataset.

The Lin CCC, coefficient of determination, and Pearson correlation coefficient are shown in Table 2. Bland-Altman plots showed a good agreement between model predictions and ground truth for the cross-validation, internal test, and NIH external test sets, though there was a slight trend toward overestimation as age increased for the cross-validation and internal test sets. Conversely, a systematic trend was not observed between predictions and ground truth in the dHCP set, as there was a trend toward underestimation as age decreased and overestimation as age increased. Model predictions had more variance with increasing age because there are fewer changes associated with myelin maturation, making it quite difficult even for experts to determine age from images, and model performance clearly reflects this clinical challenge (Figs 1, 5).

A detailed overview of all results is reported in Table 2 and shown in Figure 5.

Highest Error

Ten cases with the highest error in both the test set and the external datasets were further investigated by the pediatric radiology fellow (T.A.D.) to determine possible underlying factors. Similar factors were identified across all datasets, including motion artifacts and inaccurate or excessive skull stripping (Fig S2). In addition, two cases with some pathologic conditions were identified in the test set, a patient with hydrocephalus (eg, a case with unidentified bright objects due to neurofibromatosis type 1) and a patient with medulloblastoma.

Interrater Agreement

As readers had access to patient chronologic age, they had very good agreement in the prediction of maturation age, with an MAE of 0.34 (Fig S3).

Discussion

In this study, we found that a deep learning model ensemble of a 3D and a 2D CNN trained on brain MRI scans in 710 infants and young children up to 3 years of age could precisely predict myelin age in infancy and early childhood, with an MAE of 1.77 months (95% CI: 1.37, 2.23). Thus, our model could help to accurately predict the age of myelin and assist less experienced trainees and general radiologists who may miss cases of developmental delay.

Myelin maturation in the white matter manifests as high signal intensity on T1-weighted images and low signal intensity on T2-weighted images. The milestones of myelination differ

Table 2: Model Performance on Myelin Maturation Age Prediction

Dataset and Metrics	Models		
	2D	3D	2D + 3D
Cross–validation (*n* = 710)			
MAE	1.53 (1.41, 1.66)	2.06 (1.9, 2.24)	1.63 (1.5, 1.77)
Coefficient of determination (R^2)	0.96 (0.96, 0.97)	0.93 (0.92, 0.94)	0.96 (0.95, 0.96)
Pearson correlation coefficient (*r*)	0.98 (0.98, 0.98)	0.96 (0.96, 0.97)	0.98 (0.97, 0.98)
CCC	0.98 (0.98, 0.98)	0.96 (0.96, 0.97)	0.98 (0.97, 0.98)
Test set (*n* = 123)			
MAE	1.43 (1.08, 1.84)	2.55 (2.01, 3.17)	1.77 (1.37, 2.23)
Coefficient of determination (R^2)	0.95 (0.92, 0.98)	0.87 (0.81, 0.92)	0.93 (0.9, 0.96)
Pearson correlation coefficient (*r*)	0.98 (0.96, 0.99)	0.94 (0.91, 0.96)	0.97 (0.95, 0.98)
CCC	0.97 (0.96, 0.99)	0.93 (0.89, 0.95)	0.96 (0.94, 0.98)
dHCP (*n* = 383)			
MAE	0.44 (0.42, 0.47)	0.27 (0.25, 0.29)	0.31 (0.29, 0.33)
Coefficient of determination (R^2)	−0.95 (−1.25, −0.72)	0.17 (0.01, 0.3)	−0.04 (−0.21, 0.1)
Pearson correlation coefficient (*r*)	0.59 (0.51, 0.65)	0.67 (0.62, 0.72)	0.71 (0.66, 0.75)
CCC	0.28 (0.23, 0.33)	0.64 (0.58, 0.69)	0.52 (0.46, 0.57)
NIH (*n* = 226)			
MAE	2.26 (2.07, 2.48)	2.27 (1.98, 2.58)	1.22 (1.08, 1.37)
Coefficient of determination (R^2)	0.79 (0.73, 0.83)	0.71 (0.63, 0.78)	0.92 (0.9, 0.94)
Pearson correlation coefficient (*r*)	0.96 (0.95, 0.98)	0.96 (0.94, 0.97)	0.97 (0.96, 0.98)
CCC	0.89 (0.86, 0.91)	0.90 (0.87, 0.92)	0.96 (0.95, 0.97)

Note.—Data in parentheses are 95% CIs (nonparametric percentile bootstrapping). CCC = Lin concordance correlation coefficient, dHCP = Developing Human Connectome Project, MAE = mean absolute error, NIH = National Institutes of Health, 3D = three-dimensional, 2D = two-dimensional.

between T1-weighted and T2-weighted images: T1-weighted images are more useful at early stages, whereas T2-weighted images are more useful at late stages (3,20). Normal myelin maturation can be broadly divided into five stages (3,21). At younger than 4 months of age, an infantile pattern is observed, in which the contrast between gray and white matter is reversed compared with the adult pattern. At 5–9 months of age, the genu of the corpus callosum begins to be myelinated. At 12 months of age, the rate of myelination decreases, and a more adultlike pattern appears on T1-weighted images. Between 1 and 2 years, the adult pattern is recognized on T1-weighted images, with persistent bright signal areas on T2-weighted images. Between 2 and 3 years, terminal myelination areas manifest persistent T2 hyperintensity that is not brighter than that of the gray matter.

One of the first MRI studies on myelin maturation dates back to 1988, in which researchers evaluated the intensity changes of white matter compared with that of gray matter and found that myelin development follows a specific sequence (20). In the subsequent studies, researchers developed a simplified classification system to facilitate and standardize the assessment of myelin maturation (22,23) and generated detailed myelin-time diagrams for different anatomic locations (4). These diagrams are still used in radiology clinics to evaluate myelin maturation.

Deep learning is a subset of machine learning methods that learn from input data, without requiring engineered features to achieve the desired output, for example, image segmentation or lesion characterization (24). To date, several different deep learning models have been proposed for classification, detection, and segmentation tasks in radiology (25). Recent studies, albeit in different populations or on a small scale, show promising results that deep learning could be employed to automatically estimate brain age from MRI scans (6–11). In one of the first studies, Cole et al (6) estimated chronologic age using 3D CNNs trained on volumetric gray and white matter maps from T1-weighted brain MRI scans in 1601 healthy adults, with an MAE of 49–54 months. Shabanian et al (7) classified neurodevelopmental age into six predefined age groups with 94% accuracy using a 3D CNN architecture trained in the fusion of T1-weighted, T2-weighted, and proton density–weighted brain MRI scans in 112 infants and children up to 3 years of age. Kawaguchi et al (8) constructed a multilayer CNN with input from a single T1-weighted and T2-weighted brain MRI section at the corpus callosum level in 161 infants and children up to 2 years of age and reported an MAE of 1.9 months. Wada et al (9) employed a multilayer CNN to estimate myelin maturation in 119 children younger than 2 years of age and reported an MAE of 0.98 months. Hong et al (10) constructed a 3D CNN architecture using sagittal T1-weighted brain MRI scans in 220 healthy children younger than 5 years of age and reported an MAE of 2.3 months with data augmentation. In a recent study,

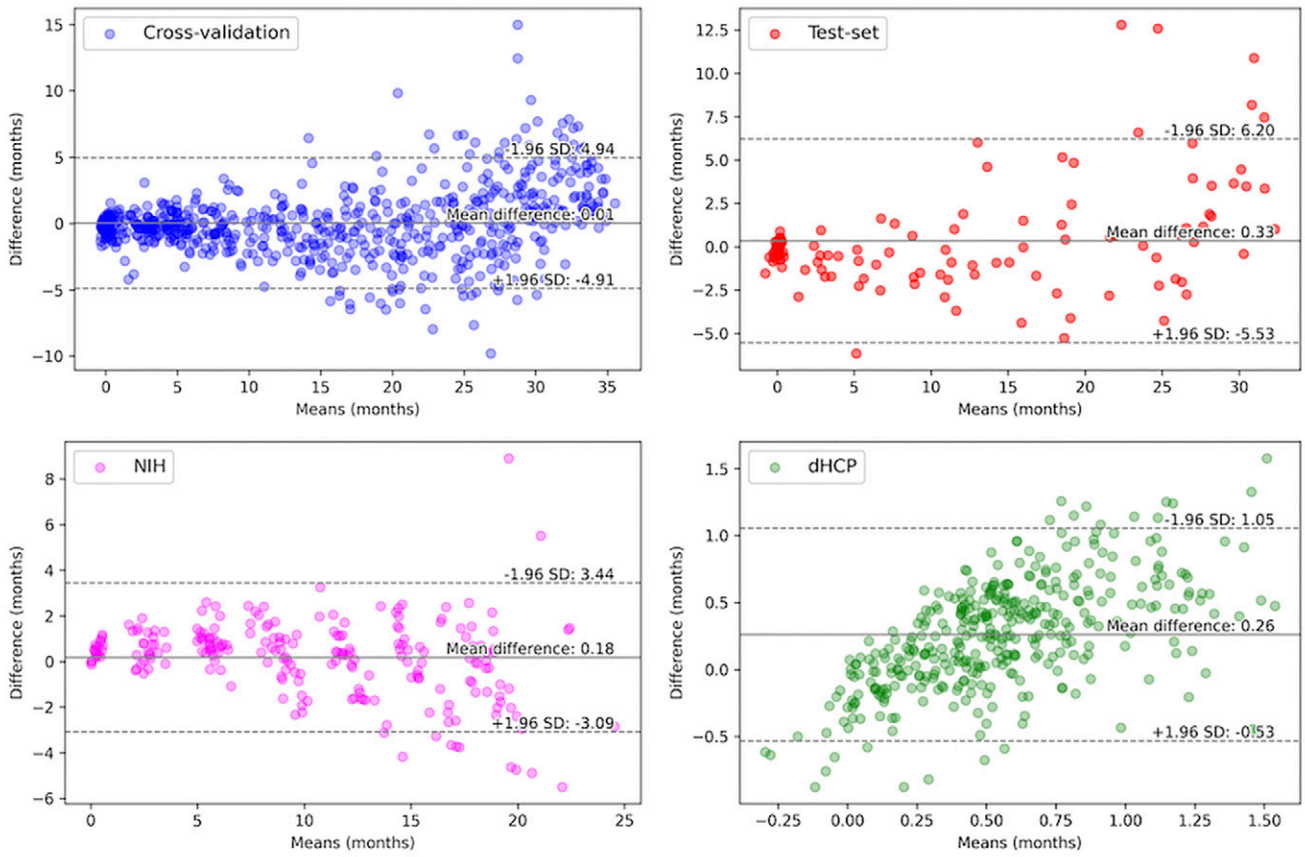


Figure 5: Bland-Altman plots of model predictions in different datasets. The difference is plotted by using the average of both the ground truth and model prediction for each patient. The center solid line represents the mean of differences. The top dashed line shows the upper limit of agreement (mean difference + 1.96 SD); the bottom dashed line shows the lower limit of agreement (mean difference – 1.96 SD). Plots show a possible relationship between ground truth versus predicted myelin maturation age. Results of the cross-validation set are depicted in blue, the internal test set in red, the National Institutes of Health (NIH) set in pink, and the Developing Human Connectome Project (dHCP) set in green.

Chen et al (11) used a multichannel 3D CNN architecture to build a brain age prediction model on a dataset with 518 studies of healthy children younger than 25 months of age and reported an MAE of 1.92 months on their test set and 2.09 months on the NIH dataset. In this study, we trained an ensemble of 3D and 2D CNNs on T1-weighted and T2-weighted brain MRI scans in 710 infants and young children up to 3 years of age. By using the same external set, our reimplementation of the 3D model of Chen et al showed similar performance (current study, MAE = 2.27 months vs Chen et al, MAE = 2.09 months). As Chen et al did not openly publish their dataset nor share it upon request, we could only reimplement their model by training it on our own dataset. Nevertheless, our ensemble model (3D + 2D) had substantially better performance (MAE = 1.22 months) on the NIH dataset and hence outperformed Chen et al's model. Our ensemble model had low error rates (MAE = 0.31 months) on the second external set, that is, the dHCP set, consisting of infants younger than 2 months of age; however, the model slightly underestimated age in younger infants and overestimated in older infants. Nevertheless, our model provides clinically relevant results given that assessment of maturation delay severity depends on the age of the patient, as in the first 0–12 months, myelin maturation accelerates at a very rapid pace and after 24 months it slows down. That is, if the patient is very young (eg, 0–2 months old), even 2 months of delay could be quite noticeable and problematic, whereas if the patient is older than 24 months, even a delay greater than 2 months could be very difficult to detect (26). Moreover, to our knowledge, our study provides the largest open-access dataset from a diverse patient sample. By making our dataset publicly available, we hope to contribute to accelerating the research in this area.

One of the limitations of our study was that we used an inhomogeneous dataset containing patients with minor pathologic conditions (eg, postoperative changes, small gliotic foci) along with completely healthy individuals. Although our model is not trained on homogeneous healthy cases as done in previous studies, we consider that our dataset represents a real-world scenario and, therefore, is more generalizable and reproducible in a clinical setting. We did not segment the images to force the model to recognize the signal intensity associated only with the myelinization, but we assume that the model focused on the manifestations due to myelin maturation anyway, as we mitigated the age-related changes in brain size by registering all images to a reference patient. Another limitation of our study was the use of different acquisition protocols over a decade, whereas the MRI scanner remained the same. Nevertheless, we mitigated these variations in the input data by image preprocessing methods. One of our external test sets, dHCP, contained only infants

younger than 2 months old and limited the model's evaluation, hence, the model constantly under- or overestimated the age. Therefore, we added a second external dataset, the NIH dataset, which contains children up to 25 months of age. At the time of the study, there were no other publicly available datasets suitable for this task. Our internal and external sets consist of only "normal cases" in terms of myelin maturation; therefore, the model has not seen cases with delayed myelin maturation, as there was no available open or private dataset at the time of the study. Further studies are needed to test the model on patients with delayed myelin maturation. In this study, we used fivefold cross-validation for hyperparameter tuning; however, it has been shown that cross-validation might have the risk of overestimating the model's performance, whereas nested cross-validation provides an unbiased estimate (19,27). Therefore, using nested cross-validation would have been a better approach; nevertheless, we think that validating our model on two external datasets balances this risk to some degree.

In conclusion, it is critical to assess myelin development to detect and treat developmental disorders in a timely manner. Determining normal myelination depends on an understanding of age-related progression, which requires time and expertise. This task may prove tedious even for radiologists with years of experience. Moreover, visual determination of myelin maturation is always participant to variability both within and between observers, as we also observed in our study. Automating the estimation of myelin maturation and implementing it into the daily workflow will save valuable time for both general and pediatric radiologists. In the long term, our method could also be used to detect pathologic conditions, such as developmental delays. Currently, there is no software or model that is compatible with PACS systems. Nevertheless, our model is openly available, and we hope that this will encourage researchers to try it out in their own institutions.

Author contributions: Guarantors of integrity of entire study, **T.A.D., A.N.D., F.P., J.W.**; 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, **T.A.D., R.A.T., A.N.D., F.P.**; clinical studies, **T.A.D., R.A.T., N.L., A.N.D., F.P.**; experimental studies, **T.A.D., R.A.T., A.N.D., F.P., J.W.**; statistical analysis, **T.A.D., A.N.D., F.P., J.W.**; and manuscript editing, **T.A.D., R.A.T., A.N.D., B.S., F.P., J.W.**

Disclosures of conflicts of interest: **T.A.D.** Member of the *Radiology: Artificial Intelligence* 2023–2025 trainee editorial board. **R.A.T.** No relevant relationships. **N.L.** No relevant relationships. **A.N.D.** Consultant and advisor for Eisai, Epilog, Idorsia, Neurocrine Biosciences, and Roche. **B.S.** No relevant relationships. **F.P.** No relevant relationships. **J.W.** No relevant relationships.

References

1. Carmody DP, Dunn SM, Boddie-Willis AS, DeMarco JK, Lewis M. A quantitative measure of myelination development in infants, using MR images. Neuroradiology 2004;46(9):781–786.
2. Parazzini C, Baldoli C, Scotti G, Triulzi F. Terminal zones of myelination: MR evaluation of children aged 20-40 months. AJNR Am J Neuroradiol 2002;23(10):1669–1673.
3. Welker KM, Patton A. Assessment of normal myelination with magnetic resonance imaging. Semin Neurol 2012;32(1):15–28.
4. Staudt M, Krägeloh-Mann I, Grodd W. Normal myelination in childhood brains using MRI--a meta analysis [in German]. Rofo 2000;172(10):802–811.
5. Matsumoto JA, Gaskin CM, Kreitel D, Kahn SL, eds. MRI Atlas of Pediatric Brain Maturation and Anatomy. New York, NY: Oxford University Press, 2015; 35–471.
6. Cole JH, Poudel RPK, Tsagkrasoulis D, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage 2017;163:115–124.
7. Shabanian M, Siddiqui A, Chen H, DeVincenzo JP. Neurodevelopmental age estimation of infants using a 3D-convolutional neural network model based on fusion MRI sequences. arXiv.2010.03963 [preprint] https://arxiv.org/abs/2010.03963. Published October 7, 2020. Accessed May 24, 2023.
8. Kawaguchi M, Kidokoro H, Ito R, et al. Age estimates from brain magnetic resonance images of children younger than two years of age using deep learning. Magn Reson Imaging 2021;79:38–44.
9. Wada A, Saito Y, Fujita S, et al. Automation of a rule-based workflow to estimate age from brain MR imaging of infants and children up to 2 years old using stacked deep learning. Magn Reson Med Sci 2023;22(1):57–66.
10. Hong J, Feng Z, Wang SH, et al. Brain age prediction of children using routine brain MR images via deep learning. Front Neurol 2020;11:584682.
11. Chen JV, Chaudhari G, Hess CP, et al. Deep learning to predict neonatal and infant brain age from myelination on brain MRI scans. Radiology 2022;305(3):678–687.
12. Cyriac J. pacs-ris-crawler/pacs-ris-crawler.github.io: PACS-RIS Crawler v1.0. 2019.. Published December 16, 2019. Accessed May 24, 2023.
13. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M. SynthStrip: skull-stripping for any brain image. Neuroimage 2022;260:119474.
14. Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can Assoc Radiol J 2019;70(2):107–118.
15. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv.1905.11946 [preprint] https://arxiv.org/abs/1905.11946. Published May 28, 2019. Accessed May 24, 2023.
16. Xie Q, Luong MT, Hovy E, Le QV. Self-Training With Noisy Student Improves ImageNet Classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; 10684–10695.
17. Falcon W, Borovec J, Wälchli A, et al. PyTorchLightning/pytorch-lightning: 0.7.6 release. Published May 15, 2020. Accessed May 24, 2023.
18. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41(4):1149–1160.
19. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. Data handling. Radiol Artif Intell 2022;4(5):e210290.
20. Barkovich AJ, Kjos BO, Jackson DE Jr, Norman D. Normal maturation of the neonatal and infant brain: MR imaging at 1.5 T. Radiology 1988;166(1 Pt 1):173–180.
21. Branson HM. Normal myelination: a practical pictorial review. Neuroimaging Clin N Am 2013;23(2):183–195.
22. Staudt M, Schropp C, Staudt F, Obletter N, Bise K, Breit A. Myelination of the brain in MRI: a staging system. Pediatr Radiol 1993;23(3):169–176.
23. Staudt M, Schropp C, Staudt F, et al. MRI assessment of myelination: an age standardization. Pediatr Radiol 1994;24(2):122–127.
24. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
25. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–510.
26. van der Knaap MS, Valk J, eds. Myelin and White Matter. In: Magnetic Resonance of Myelination and Myelin Disorders. Berlin, Germany: Springer Berlin Heidelberg, 2005; 1–19.
27. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7(1):91.