Neonatology

Global Neonatal Research

Neonatology 2022;119:311–319
DOI: 10.1159/000522629

Received: January 5, 2022
Accepted: February 12, 2022
Published online: March 31, 2022

Outcome Prediction and Inter-Rater Comparison of Four Brain Magnetic Resonance Imaging Scoring Systems of Infants with Perinatal Asphyxia and Therapeutic Hypothermia

Juliette F. Langeslag[a, b] Floris Groenendaal[c] Stefan D. Roosendaal[d] Linda S. de Vries[c] Wes Onland[a] Mariska M.G. Leeflang[e] Paul F.C. Groot[d] Anton H. van Kaam[a, b] Timo R. de Haan[a] on behalf of the PharmaCool Study Group

[a]Department of Neonatology, Emma Children's Hospital, Amsterdam University Medical Centers, Amsterdam, The Netherlands; [b]Amsterdam Reproduction & Development Research Institute, Amsterdam University Medical Centers, Amsterdam, The Netherlands; [c]Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, and Brain Center, Utrecht, The Netherlands; [d]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Amsterdam, The Netherlands; [e]Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

Keywords
Perinatale asphyxia · Fullterm neonate · Neurodevelopmental outcome · Prediction · Magnetic resonance imaging scoring

Abstract
Introduction: The brain magnetic resonance imaging (MRI) result is a major predictor for the outcome of term infants with perinatal asphyxia who underwent therapeutic hypothermia. In daily practice, no uniform method is used to assess these images. ***Purpose:*** The aim of this study was to determine which MRI-score best predicts adverse outcome at 24 months of age and has the highest inter-rater reliability. ***Methods:*** Four MRI scoring systems for term infants with perinatal asphyxia were selected: Rutherford score, Trivedi score, Weeke score, and NICHD NRN score. Experienced blinded raters retrospectively evaluated the brain MR Images of 161 infants using all four scoring systems. Long-term outcome (the composite outcome death or adverse outcome, and its separate components) were routinely assessed by standardized testing at the age of 24 months. The predictive accuracy was assessed by logistic regression analyses and expressed as area under the ROC curve (AUC). The inter-rater reliability of the scores was calculated by the weighted Kappa or intraclass correlation. A sensitivity analysis using only high-quality MRI scans was performed. ***Results:*** All four MRI scoring systems demonstrated an AUC of >0.66 for the prediction of adverse outcome and ≥0.80 for the prediction of death. The inter-rater reliability analyses demonstrated the highest reliability for the Weeke and Trivedi scores. When only assessing the high-quality scans, the AUC increased further. ***Conclusion:*** All four MRI brain scores proved reliable predictors for an adverse outcome at 24 months of age. The Weeke and Trivedi score demonstrated the highest inter-rater reliability. The use of high-quality MRI further improved prediction.

A complete list of study group members appears in the Acknowledgments.

Karger@karger.com
www.karger.com/neo

Correspondence to:
Timo Robert de Haan, t.r.dehaan@amsterdamumc.nl

Introduction

Predicting neurodevelopmental outcome of term infants with hypoxic-ischemic encephalopathy following perinatal asphyxia and therapeutic hypothermia (TH) remains challenging. Magnetic resonance imaging (MRI) of the neonatal brain is an important diagnostic and prognostic tool in these patients especially after the introduction of diffusion-weighted imaging (DWI), and MR spectroscopy (1H-MRS). For optimal detection of ischemia with DWI, the brain of asphyxiated infants should be scanned in the first 8 days of life [1–3].

Despite the important role of neonatal brain MRI in predicting long-term outcome after asphyxia and hypothermia, no uniform method of assessing these MR images is currently used. In daily practice, the prognosis is made without the use of an MRI scoring system. An objective and reliable MRI scoring system might provide more uniformity in imaging assessments, improve the quality of multidisciplinary discussions on prognosis, and improve benchmarking.

Multiple MRI scoring systems have been described in the literature but to date no comparison has been performed in a large cohort of asphyxiated newborns treated with TH. It is unknown which MRI scoring system performs best in predicting outcome and should be introduced in clinical practice.

The main objective of this study was to assess and compare the predictive value of four clinical currently used MRI scoring systems when applied to MR images of a large multicenter cohort of asphyxiated term infants treated with therapeutically hypothermia [4]. The secondary objective was to compare the inter-rater reliability of these selected MRI scoring systems.

Methods

Study Design and Participants

The PharmaCool study was a multicenter prospective observational cohort study performed between November 2010 and October 2014 in 11 neonatal intensive care units in the Netherlands and Belgium. It investigated how TH after asphyxia influenced the pharmacokinetics and dynamics of administered drugs. The Institutional Review Board of each participating center approved the study [4]. The current MRI study included all infants of the PharmaCool study. Infants with missing or insufficient quality of MRI images (e.g., due to movement artifacts or technical-imaging protocol errors), who were lost to follow-up or had congenital abnormalities of the brain were excluded.

MRI Procedures during Initial Hospitalization

The MRI protocol and hardware specifications of each participating center are shown in online suppl. Table A1 (for all online suppl. material, see www.karger.com/doi/10.1159/000522629). In the majority of infants, the MR images were collected within the first week of life and after 72 h of controlled hypothermia treatment. They consisted of T1- and T2-weighted images, DWI or Diffusion Tensor Imaging, and 1H-MRS if available. In 7/11 centers, the MR images were collected without sedation; immobilization was achieved by vacuum matrasses or by a specialized neonatal MRI incubator.

Study Procedures and Outcomes

Description of MRI Scoring Systems

We identified 11 MRI scoring systems [5–15] used for perinatal asphyxia and selected four of them: the Rutherford score [8], the Trivedi score (adapted Bednarak score) [14], the Weeke score [15], and the NICHD score (National Institute of Child Health and Human Development) [16]. The following criteria were used for selection: year of publication, citation number, newly published update of previous scoring system, anatomical correctness, the extent of brain injury pattern description matching with the hypoxic-ischemic encephalopathy brain injury patterns, the imaging sequences used and clinical applicability. Online suppl. Table A2 shows further details of these scoring systems.

Conceptually, the four selected scoring systems were comparable, with a higher score indicating a more severe pattern of brain damage. However, the scoring systems differed in structure or sub-scoring items. Details for each of the four MRI scoring systems are depicted in online suppl. Table A2.

The quality of the MRIs was scored as high, moderate, poor, or unable to assess. Scan quality was visually assessed based on: slice thickness, the magnet field strength (Tesla), contrast of the images, and artifacts.

Scoring the MR Images

The MR images were rated by highly experienced experts at two study locations in the Netherlands (at the UMCU by LdV and FG; at the UMCA by SR). To calculate the predictive value of each scoring system, study location UMCU all included MR images with all four scoring systems. To assess the inter-rater reliability of each scoring system, study location UMCA scored the MR images in 100 randomly selected infants for comparison.

Raters were blinded for the center of origin, the clinical course, and outcome. Raters did receive information on gestational age and postnatal age at MRI examination,

necessary for the correct interpretation of brain imaging (e.g., development, myelination, DWI, and Diffusion Tensor Imaging).

Outcome Definitions

All infants attended standardized outpatient follow-up visits at 24 months of age in all centers. Neurodevelopmental outcome was assessed with the Bayley Scales of Infant and Toddler Development (3rd edition, Dutch language [BSID-III-NL]). If applicable, the level of cerebral palsy was classified using the Gross Motor Function Classification System [17].

Adverse outcome in survivors was defined as a test score of ≥1 standard deviation below the reference mean on the BSID-III-NL composite cognitive score or composite motor score (e.g., a score <85 points) or a Gross Motor Function Classification System of ≥2, or hearing loss requiring hearing aids or severe visual impairment (blind of abnormal vision) at 24 months. For the statistical analyses, the binary outcomes of interest at 24 months were the composite outcome death or adverse outcome, and its separate components.

Bayley Scales of Infant and Toddler Development: Corrective Measures

Almost all participating centers used the BSID-III-NL for assessing neurodevelopmental outcome, but for 3 infants the second edition of the Dutch BSID and for 49 infants the American norms for the BSID were used. To compensate for discrepancies between these scales, an evidence-based method of correction was performed as previously published [18].

Statistical Analysis

Data were analyzed in R statistical software (Version 3.6.3 for Windows) and R Studio (integrated development for R, Boston, 2015) (R studio desktop 1.2.5033). Descriptive statistics summarized patient characteristics and outcome parameters. We reported normally distributed data as mean with standard deviation and non-normally distributed data as median with interquartile range (IQR).

To retrieve missing data, the first author contacted the principal investigators of each participating center. If clinical data could not be retrieved and there was <50% missing data, multiple imputation was performed on the data set. For the dependent variables, passive imputation was used. We checked possible correlations for each variable. Multiple imputation results were checked for imputation errors, with convergence plots, strip plots, and checking the individual datasets [19].

To investigate the predictive value of each MRI scoring system for the outcomes of interest, multivariable logistic regression models were constructed for each outcome. First, we assessed the predictive value of the MRI scoring systems as a sole predictor of the outcome by calculating the area under the receiver operator characteristics curve (AUC). Next, we assessed the predictive value of a model containing the following known variables impacting the outcomes of interest: birthweight, Thompson score, amplitude-integrated electroencephalography (EEG) confirmed seizures, and baseline pH (i.e., umbilical cord pH or first available pH in hospital). The MRI scoring systems were added individually to the initial model and the AUCs were calculated. We compared these AUCs by calculating *Z*-statistics, as described by DeLong et al. [20], *Z*-statistics are used when comparing AUCs derived from the same sample of infants. The larger the correlation between the diagnostic algorithm the more sensitive the paired *Z*-test will be. The cutoff of a *Z*-statistic can then be established using a normal distribution table [20].

The inter-rater reliability was assessed using the weighted Cohen's kappa coefficient (wKappa) or intraclass correlation coefficient (ICC) depending on whether the scoring system was numeric or categorical [21, 22]. We interpreted the kappa values following Landis et al. [23], *a* value >0.6 indicated substantial agreement, *a* value >0.8 indicated almost perfect agreement compared to chance variation.

For the ICC, the two-way random effect models and "single rater" unit were used. ICC values between 0.5 and 0.75 were interpreted as moderate agreement and between 0.75 and 0.9 as good agreement [24].

To test the robustness of the results, sensitivity analyses were performed. We performed a sensitivity analysis for both 24 months outcome and the inter-rater reliability using only the highest quality scans.

Furthermore, we retested the predictive value of the Rutherford score after including the DWI sequence, which was not part of the original scoring system. This analysis was only performed for the MR images of infants below 8 days after birth when pseudo-normalization has not yet occurred.

Results

Patient Flow Diagram and Characteristics

In total, 189 infants were enrolled in the PharmaCool study. We excluded 15 (7.9%) infants and 13 (7.5%) infants were lost to follow-up at 24 months of age, leaving

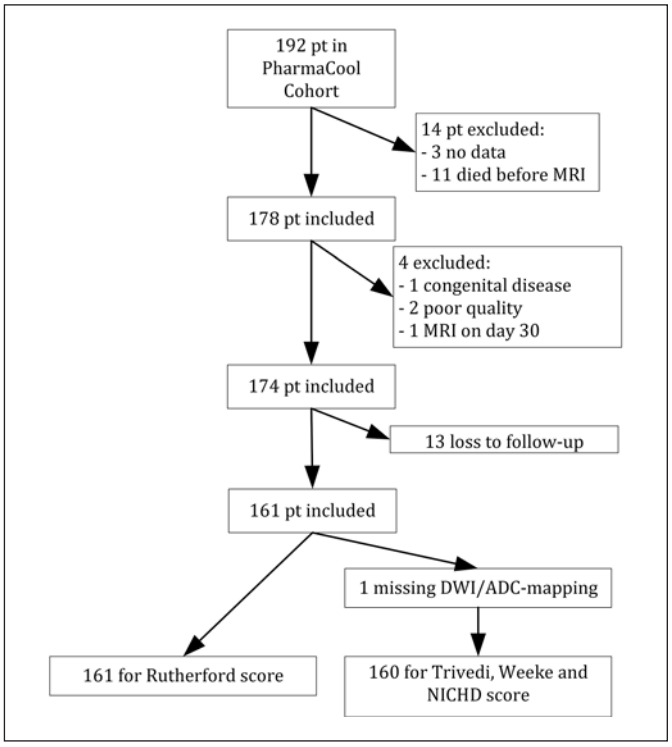

Fig. 1. Patient flowchart.

Table 1. Patient characteristics

	Total (*N* = 161)
Sex (male), *n* (%)	95 (59)
Gestational age,* weeks	40.0 (3)
Birthweight,* g	3,380 (780)
Apgar at 5 min*	3 (2)
Missing, *n* (%)	5 (3.1)
pH (first)*	6.98 (0.26)
Missing, *n* (%)	1 (0.6)
Thompson score*	9 (3)
Missing, *n* (%)	14 (8.7)
aEEG proven seizures, *n* (%)	56 (34.8)
Missing, *n* (%)	11 (6.8)
GMFCS,»*n* (%)	12 (9.1)
Vision,»*n* (%)	3 (2.3)
Hearing,»*n* (%)	2 (1.5)
CCS»,*	101 (19)
CMS»,*	104 (18)
Death, *n* (%)	29 (18.0)
Adverse outcome,»*n* (%)	28 (21.2)
Missing, *n* (%)	25 (18.9)
Combined outcome, *n* (%)	57 (35.4)
Missing, *n* (%)	25 (15.5)
Month of follow-up*	24 (1)
Normal outcome,» *n* (%)	79 (59.8)
Day of MRI*	6 (2)
MRI <8 days, *n* (%)	134 (83.2)

Nonmissing data. GMFCS, the Gross Motor Function Classification System (at 24 months); CCS, composite cognitive score (BSID-III); CMS, composite motor score (BSID-III). * Median (IQR). » Total = 132.

161 patients for analyses (Fig. 1). Table 1 shows the patient characteristics of this cohort.

The mortality during hospital stay was 29/161 (18%). Adverse outcome was present in 28/132 (21%) and normal outcome in 79/132 (60%). Outcome assessment was incomplete in 25/132 (19%) of the survivors.

MRI Scoring Systems and Long-Term Outcome

Table 2 shows the results of the univariable logistic regression analysis for the prediction of outcome at 24 months. Since 1H-MRS could only be retrieved in 2/11 centers, the Weeke score was assessed without 1H-MRS. All MRI scoring systems had an acceptable prediction of adverse outcome with comparable AUCs between 0.66 and 0.71. There was an excellent prediction of death for most scoring systems with AUCs ≥0.90, except for the Rutherford score (AUC 0.84). A good prediction was observed for the composite outcome death or adverse, with all scoring systems reaching an AUC close to or above 0.80. Only the NICHD performed significantly better than the Rutherford score for the composite outcome.

Table 3 shows the effect of adding the different MRI scoring systems to the initial model incorporating known clinical predictors for the different outcomes at 24 months. Adding the different MRI scoring systems resulted in a (modest) improvement of the AUC for the adverse outcome, with only the NICHD score reaching statistical significance when compared to the initial model. Adding the analyses results of the MRI scoring systems significantly improved the prediction of death; composite outcome or adverse outcome with all scoring systems reaching an AUC close to or above 0.90. Only the Trivedi score did not perform significantly better than the initial model. A plot plotting the raw MRI score points versus the Bayley-III-NL composite cognitive score points can be found in online suppl. Figure A2.

Inter-Rater Reliability

From the 100 randomly selected patients, two MRIs were excluded: one because of an unexpected existing congenital abnormality and one because the MRI scan was performed post-mortem. As a result,

Table 2. The prediction of the outcomes, univariate logistic regression

	Univariate logistic regression								
	adverse			death			composite		
	AUC	*z*-statistic	*p* value	AUC	*z*-statistic	*p* value	AUC	*z*-statistic	*p* value
Rutherford	0.66 (0.54–0.76)	*	*	0.84 (0.72–0.91)	*	*	0.79 (0.70–0.86)	*	*
Trivedi	0.66 (0.55–0.77)	−0.22	0.79	0.96 (0.92–0.98)	−2.36	0.02	0.85 (0.77–0.90)	−1.22	0.22
Weeke	0.71 (0.58–0.81)	−1.32	0.21	0.96 (0.91–0.98)	−2.35	0.02	0.88 (0.80–0.93)	−1.82	0.07
NICHD	0.69 (0.58–0.78)	−0.69	0.51	0.95 (0.91–0.98)	−2.12	0.03	0.86 (0.78–0.91)	−1.95	0.06

AUC, area under the receiver operating curve. * Reference.

Table 3. The prediction of the outcomes, multiple logistic regression

	Multiple logistic regression								
	adverse			death			composite		
	AUC	*z*-statistic	*p* value	AUC	*z*-statistic	*p* value	AUC	*z*-statistic	*p* value
Model	0.68 (0.54–0.78)	*	*	0.85 (0.74–0.91)	*	*	0.81 (0.72–0.88)	*	*
Rutherford	0.73 (0.60–0.82)	−1.20	0.25	0.91 (0.83–0.96)	−2.11	0.04	0.87 (0.80–0.92)	−2.14	0.03
Trivedi	0.72 (0.60–0.81)	−1.20	0.26	0.97 (0.92–0.99)	−2.82	0.01	0.89 (0.82–0.93)	−1.70	0.10
Weeke	0.75 (0.62–0.84)	−1.65	0.12	0.97 (0.93–0.99)	−3.01	0.003	0.91 (0.84–0.95)	−2.30	0.03
NICHD	0.76 (0.63–0.85)	−2.03	0.05	0.97 (0.93–0.99)	−3.33	0.001	0.92 (0.85–0.96)	−3.32	0.001

AUC, area under the receiver operating curve (with 95% CI); parameters used in multiple logistic regression: birthweight, pH (first), Thompson score, seizures (during hospital stay) and MRI score. * Reference.

Table 4. The inter-rater reliability

Score	Total (*N* = 98), ICC (95% CI)
Rutherford score	0.53 (0.34–0.66)
Trivedi score	0.76 (0.66–0.83)
Weeke score	0.74 (0.64–0.82)
NICHD score*	0.65 (0.51–0.78)

ICC, intraclass correlation. * Weighted kappa.

both study locations scored a total of 98 MRI scans with all four scoring systems. As shown by Table 4, the Trivedi score and Weeke score had the highest inter-rater reliability.

Sensitivity Analyses

For 84 patients (3 centers) with a high-quality MRI scan result, the analyses were repeated. Figure 2 illustrates scan quality difference. As shown by online suppl. Table A3, A4, the AUC of each scoring increased only when assessing these high-quality scans. This was seen in the univariable logistic regression as well as the multivariable logistic regression analysis.

Higher AUCs were observed when the DWI was taken into account for the Rutherford score compared with their original scoring system (online suppl. Table A5). With high-quality MRI-scans analyses, the inter-rater reliability improved for each scoring system. An example of the added value of DWI is shown in online suppl. Figure A1.

Discussion

No statistically significant differences were found between the four assessed MRI scoring systems in their discriminating power to predict the composite outcome and its separate component "adverse outcome" at 24 months.

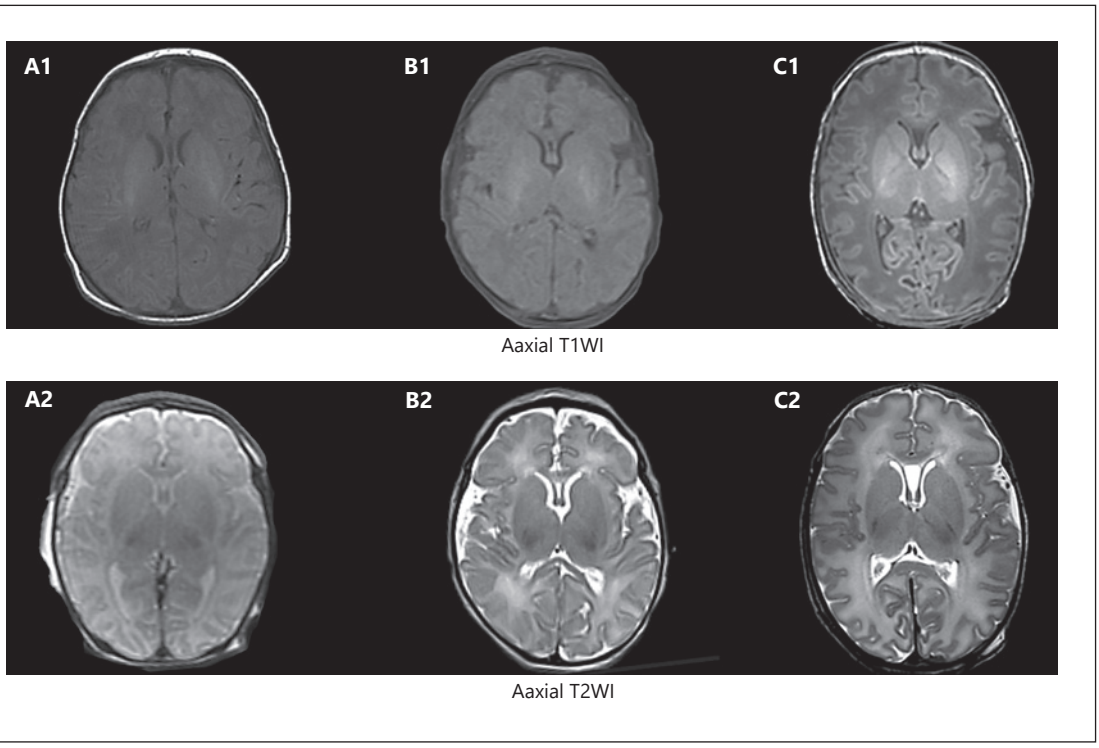

Fig. 2. Differences in quality of the MRI, the T1-and T2-weighted images of 3 patients in 3 different centers. Clear differences in contrast and detail. All images performed during first week of life. Patient A and B: 1.5 T (4 and 2 mm slice thickness resp.), patient C: 3 T (2 mm slice thickness).

Furthermore, our study shows that the Trivedi score and the Weeke score had the highest inter-rater reliability.

The observed mortality (20%) in our cohort was lower than described during the introduction of TH [25]. That the mortality is lower than before the introduction of TH is in accordance with the recent studies on outcome [26].

The AUCs for the prediction of death were higher than the AUCs for adverse outcome for each scoring system. This might be explained by redirection of care when severe brain injury was diagnosed on MRI in combination with other investigations (especially clinical exam, the amplitude-integrated EEG and EEG). The Rutherford score has a significantly lower discriminating power to predict the outcome of death than the other scoring systems, when the score was performed without the use of DWI.

The prediction of adverse outcome by the Trivedi score approximates the predictive value as described in the original article [14]. The AUC of the Weeke score in this study is comparable to Swedish cohort of the original article [15] but lower than the cohort from Utrecht. An explanation may be that the quality of MR images differed due to different imaging protocols used in the participating centers. Our finding that the AUC for all outcomes improved when applying the Weeke score only to high-quality scans seems to support this explanation. In addition, the original Weeke score included 1H-MRS, one of the best predictors of an adverse outcome after asphyxia and hypothermia [1, 3].

We were not able to compare our results statistically with the original publications of the NICHD and Rutherford et al. [8], as both publications described their predictive value as sensitivity, specificity, positive predictive value, and negative predictive values [8, 16].

When selecting a reliable predictive tool for daily practice to predict outcome at 24 months independently from the assessor, inter-rater reliability is a factor to be taken into account. Our study demonstrated that the Trivedi and Weeke scoring systems performed best in the inter-rater reliability. Even when assessing only the best quality MRI's, the Rutherford score is still much lower than the other scores, which may have been caused by the fact that the Rutherford score does not take into account the DWI. As also other important clinical variables are important

for outcome prediction, we also performed a multivariate analysis taking these into account.

The results showed that adding the MRI scoring system to these clinical characteristics significantly improved the prediction of the composite outcome and death. For the adverse outcome, only the AUC of the NICHD score improved significantly.

The strength of our study is that we assessed four peer-reviewed MRI scoring systems in a large multicenter cohort of asphyxiated infants undergoing TH. A blinded assessment was performed, both by the neonatologists as by a pediatric neuroradiologist.

A limitation of this study was the need to use multiple imputation for missing data, and the need to convert the BSID scores in a minority of infants. Although we used previously published conversion methods, we cannot exclude that over- or underestimation of the outcome results occurred.

Another limitation of our study is the variable quality of MR images which was due to heterogeneity in the MRI protocols, differences in scanner brands and field strength. On the other hand, this mirrors normal daily clinical care and therefore improves the generalizability of our results. Analyzing only high-quality MRI scans, led to improvement of the AUCs as well the inter-rater reliability for all scoring systems. This emphasizes the importance of high-quality MRI in this population.

Unfortunately, the added value of performing and rating 1H-MRS could not be assessed in this study as 1H-MRS was not performed as a standard of care at the time of study inclusion. 1H-MRS is an important contributor in diagnostic neonatal MRI [27]. Although it is always difficult to draw definitive conclusions, the current study suggests that based on prediction, inter-rater reliability and the ease of using the scoring system, the MRI scores can be rated as follows: (1) Weeke, (2) Trivedi, (3) NICHD, and (4) Rutherford [28].

As the number of children with an adverse outcome at 5.5 years could differ from 24 months – a phenomenon known as "growing into deficit," neurodevelopmental outcome should be assessed at different time intervals during follow-up. For this reason, we are currently analyzing the 5.5 year follow-up data and will also see children again at 8 year.

Conclusion

Performance was similar for all four assessed MRI scoring systems in the prediction of death, adverse and composite outcome at 24 months for term infants with perinatal asphyxia and treated with hypothermia. Inter-rater reliability was best for the Trivedi and the Weeke score. High-quality MRI using standardized protocols is essential and improves the prediction of outcome at 24 months.

Acknowledgments

PharmaCool study group members: Chris H. P. van den Akker, MD, PhD (Department of Neonatology, Emma Children's Hospital, Amsterdam UMC, location VUMC, Amsterdam, The Netherlands), Willem P. de Boode, MD, PhD (Department of Neonatology, Radboud University Medical Center-Amalia Childrens Hospital, Nijmegen, The Netherlands), Filip Cools, prof (Department of Neonatology, Universitair Ziekenhuis Brussel, Brussel, Belgium), Peter H. Dijk, MD, PhD (Department of Neonatology, University Medical Center Groningen, Beatrix Children's Hospital, University of Groningen, Groningen, The Netherlands), Koen P. Dijkman, MD (Department of Neonatology, Maxima Medical Center Veldhoven, The Netherlands), Floris Groenendaal, MD, PhD (Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands), Timo R. de Haan, MD, PhD (Department of Neonatology, Emma Children's Hospital, Amsterdam UMC, location AMC, Amsterdam, The Netherlands), Sinno H. P. Simons, MD, PhD (Department of Neonatology, Sophia Children's Hospital, Erasmus MC, Rotterdam, The Netherlands), Sylke J. Steggerda, MD, PhD (Department of Neonatology, Leiden University Medical Center, Leiden, The Netherlands), Henrica L. M. van Straaten, MD, PhD (Department of Neonatology, Isala Medical Center, Zwolle, The Netherlands), and Alexandra Zecic, MD (Department of Neonatology, Universitair Ziekenhuis Gent, Gent, Belgium) are the local investigators at the participating centers and made substantial contributions to the concept and design of the study, and interpretation of data.

Statement of Ethics

This study protocol was reviewed and approved by Medic Ethic Review Committee (METC) Amsterdam UMC location AMC W20_533. Written informed consent was obtained from participants (or their parent/legal guardian/next of kin) to participate in the study.

Conflict of Interest Statement

Dr. de Haan reports grants from The Netherlands Organization for Health Research and Development ZonMW during the conduct of the study. No other disclosures were reported.

Funding Sources

This study was funded by a Project Grant from The Netherlands Organization for Health Research and Development ZonMW Priority Medicines for Children Grant No.: 40-41500-98-

9002. Role of Funder/Sponsor: the funding agency had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Author Contributions

Juliette Langeslag prepared database, performed the statistical analyses, prepared the data tables, drafted the initial manuscript, and revised the manuscript. Linda de Vries, Floris Groenendaal, Stefan Roosendaal blinded scored MR images and reviewed the manuscript for important intellectual content. Mariska Leeflang made substantial contributions to the interpretation of data and reviewed and revised the manuscript. Paul Groot made substantial contributions to the MRI data management and archiving. Timo de Haan, Wes Onland, and Anton van Kaam are local investigators, made substantial contributions to the concept and design of the study, had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis, and critically reviewed the manuscript for important intellectual content. All the authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Data Availability Statement

Data generated or analyzed during the study are available from the corresponding author by request.

References

1 Alderliesten T, de Vries LS, Benders MJ, Koopman C, Groenendaal F. MR imaging and outcome of term neonates with perinatal asphyxia: value of diffusion-weighted MR imaging and (1)H MR spectroscopy. Radiology. 2011 Oct;261(1):235–42.
2 Groenendaal F, de Vries LS. Fifty years of brain imaging in neonatal encephalopathy following perinatal asphyxia. Pediatr Res. 2017;81(1–2):150–5.
3 Wisnowski JL, Wintermark P, Bonifacio SL, Smyser CD, Barkovich AJ, Edwards AD, et al. Neuroimaging in the term newborn with neonatal encephalopathy. Semin Fetal Neonatal Med. 2021;26(5):101304.
4 de Haan TR, Bijleveld YA, van der Lee JH, Groenendaal F, van den Broek MPH, Rademaker CMA, et al. Pharmacokinetics and pharmacodynamics of medication in asphyxiated newborns during controlled hypothermia. The PharmaCool multicenter study. BMC pediatrics. 2012;12:45.
5 Barkovich AJ, Hajnal BL, Vigneron D, Sola A, Partridge JC, Allen F, et al. Prediction of neuromotor outcome in perinatal asphyxia: evaluation of MR scoring systems. AJNR Am J Neuroradiol. 1998 Jan;19(1):143–9.
6 Haataja L, Mercuri E, Guzzetta A, Rutherford M, Counsell S, Flavia Frisone M, et al. Neurologic examination in infants with hypoxic-ischemic encephalopathy at age 9 to 14 months: use of optimality scores and correlation with magnetic resonance imaging findings. J Pediatr. 2001 Mar;138(3):332–7.
7 Jyoti R, O'Neil R, Hurrion E. Predicting outcome in term neonates with hypoxic-ischaemic encephalopathy using simplified MR criteria. Pediatr Radiol. 2006;36(1):38–42.
8 Rutherford M, Ramenghi LA, Edwards AD, Brocklehurst P, Halliday H, Levene M, et al. Assessment of brain tissue injury after moderate hypothermia in neonates with hypoxic-ischaemic encephalopathy: a nested substudy of a randomised controlled trial. Lancet Neurol. 2010 Jan;9(1):39–45.
9 van Rooij LG, Toet MC, van Huffelen AC, Groenendaal F, Laan W, Zecic A, et al. Effect of treatment of subclinical neonatal seizures detected with aEEG: randomized, controlled trial. Pediatrics. 2010 Feb;125(2):e358–66.
10 Bonifacio SL, Glass HC, Vanderpluym J, Agrawal AT, Xu D, Barkovich AJ, et al. Perinatal events and early magnetic resonance imaging in therapeutic hypothermia. J Pediatr. 2011;158(3):360–5.
11 Bednarek N, Mathur A, Inder T, Wilkinson J, Neil J, Shimony J. Impact of therapeutic hypothermia on MRI diffusion changes in neonatal encephalopathy. Neurology. 2012 May 1;78(18):1420–7.
12 Cheong JL, Coleman L, Hunt RW, Lee KJ, Doyle LW, Inder TE, et al. Prognostic utility of magnetic resonance imaging in neonatal hypoxic-ischemic encephalopathy: substudy of a randomized trial. Arch Pediatr Adolesc Med. 2012 Jul 1;166(7):634–40.
13 Rollins N, Booth T, Morriss MC, Sanchez P, Heyne R, Chalak L. Predictive value of neonatal MRI showing no or minor degrees of brain injury after hypothermia. Pediatr Neurol. 2014 May;50(5):447–51.
14 Trivedi SB, Vesoulis ZA, Rao R, Liao SM, Shimony JS, McKinstry RC, et al. A validated clinical MRI injury scoring system in neonatal hypoxic-ischemic encephalopathy. Pediatr Radiol. 2017 Oct;47(11):1491–9.
15 Weeke LC, Groenendaal F, Mudigonda K, Blennow M, Lequin MH, Meiners LC, et al. A novel magnetic resonance imaging score predicts neurodevelopmental outcome after perinatal asphyxia and therapeutic hypothermia. J Pediatr. 2018 Jan;192:33–40.e2.
16 Laptook AR, Shankaran S, Barnes P, Rollins N, Do BT, Parikh NA, 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 Mar;230:106–11.e6.
17 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 Apr;39(4):214–23.
18 Lees CC, Marlow N, van Wassenaer-Leemhuis A, Arabin B, Bilardo CM, Brezinka C, et al. 2 year neurodevelopmental and intermediate perinatal outcomes in infants with very preterm fetal growth restriction (TRUFFLE): a randomised trial. Lancet. 2015 May 30;385(9983):2162–72.
19 Buuren SV. Flexible imputation of missing data. 2nd ed. Chapman and Hall/CRC; 2018.
20 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.
21 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968 Oct;70(4):213–20.
22 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Method. 1996;1(1):30–46.

23 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.
24 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016 Jun;15(2):155–63.
25 Jacobs SE, Berg M, Hunt R, Tarnow-Mordi WO, Inder TE, Davis PG. Cooling for newborns with hypoxic ischaemic encephalopathy. Cochrane Database Syst Rev. 2013 Jan 31(1):Cd003311.
26 Hage L, Jeyakumaran D, Dorling J, Ojha S, Sharkey D, Longford N, et al. Changing clinical characteristics of infants treated for hypoxic-ischaemic encephalopathy in England, Wales and Scotland: a population-based study using the National Neonatal Research Database. Arch Dis Childhood Fetal Neonatal Ed. 2021 Sep;106(5):501–8.
27 Lally PJ, Montaldo P, Oliveira V, Soe A, Swamy R, Bassett P, et al. Magnetic resonance spectroscopy assessment of brain injury after moderate hypothermia in neonatal encephalopathy: a prospective multicentre cohort study. Lancet Neurol. 2019 Jan;18(1):35–45.
28 Ní Bhroin M, Kelly L, Sweetman D, Aslam S, O'Dea MI, Hurley T, et al. Relationship between MRI scoring systems and neurodevelopmental outcome at two years in infants with neonatal encephalopathy. Pediatr Neurol. 2021 Oct 13;126:35–42.