Official Title: HIGH DOSE ERYTHROPOIETIN FOR ASPHYXIA AND ENCEPHALOPATHY

NCT number: NCT02811263

Document: Statistical Analysis Plan (SAP) version 2.0

Document Date: February 15, 2022

A randomized, placebo-controlled, double-masked 500 - subject clinical trial of erythropoietin for the treatment of neonatal hypoxic-ischemic encephalopathy (HIE)

Statistical Analysis Plan

TABLE OF CONTENTS

1. General Design Considerations

1.1 Hypothesis

Epo given to infants with moderate-severe HIE will safely lead to improved neurodevelopmental outcome at 24 months.

1.2 Design

This study is a randomized parallel group double-masked, placebo-controlled trial in neonates with HIE. We will enroll 500/2 = 250 subjects in each of two treatment groups: Epo-treated vs. placebo control. Eligible neonates will be enrolled and treated with either Epo or normal saline (on study day 1,2,3,4, and 7) and followed for a fixed period of 24 months (±56 days) for survival and neurological outcomes. For extenuating circumstances, for example, COVID-19 restrictions, the in-person elements of the final endpoint evaluation may be performed up to 36 months of age. The follow-up period was chosen to provide meaningful motor and cognitive outcomes, since CP and motor and cognitive deficits are more reliably diagnosed at 2 years of age. Enrolled children will be randomized, stratified by study site and severity of HIE, with equal likelihood to receive Epo or saline.

1.3 Randomization

Randomization sequences will be created centrally by the Data Coordinating Center (DCC). The DCC will provide each site pharmacy with a randomization look-up table that pairs each study ID with a treatment allocation. Randomization is stratified by site and HIE severity level ("moderate" vs. "severe" HIE) based on the modified Sarnat score. The site pharmacy will then draw up the assigned study drug dose into a syringe, based on the randomization list. The appearance of the syringes of Epo and placebo are identical, thus the bedside staff remain masked. Only the research pharmacist and DCC biostatistician will be un-masked to treatment assignment. All parents, study and clinical staff, as well as all Clinical Coordinating Center (CCC) members, will remain masked to treatment assignment throughout follow-up. To ensure balance within clinical centers, the randomization will be implemented using a randomly permuted blocks design to ensure approximate balanced assignments within site and HIE severity level. The sequence of assignments for each site and stratum will be prepared in advance by the DCC.

1.4 Analysis

Analyses will be based on a modified Intention to Treat (mITT) approach. In this approach, all randomized neonates who received at least one dose of study drug will be included in the analyses. Neonates whose parents withdrew consent after randomization and before the first dose will be excluded from the mITT.

2. Primary Outcome

We used standardized, validated neurological and developmental assessments at 24 months:

- CP diagnosed by Standardized Neurological Examination[1,2]
- Bayley III Cognitive Score
- GMFCS[3]

The **primary outcome is the composite of death or any neurodevelopmental impairment**. Neurodevelopmental impairment (mild, moderate, or severe) is defined as any of the following:

- GMFCS level ≥1, or

- GMFCS = 0 or 0.5 AND CP (any type), or
- Bayley III Cognitive Score < 90

Due to COVID-19 restrictions, the Bayley III, Neuro examination, and GMFCS data collection windows were extended from their original window of 22-26 months of age up to 36 months of age. The Bayley III is normed by age through 36-months, and CP status is not anticipated to change between 22 and 36 months of age. Therefore, this window extension does not affect the validity of the primary outcome measure.

2.1 Motor Deficit - Cerebral Palsy (CP)

Cerebral palsy (CP) will be determined by a Standardized Neurologic Examination[1,2] conducted under the direction of Dr. Karl Kuban (Co-I), using the systematized exam and video-based certification system created for the NINDS-funded ELGAN and PENUT studies.[1,2] This training program provides a formal method of neurologic testing that is highly reproducible. The exam was created specifically to determine the presence and classification of CP (i.e., quadriplegic-QP, hemiplegic-HP and diplegic-DP) at age 2. To standardize the quality of neurologic examination findings, two follow-up examiners at each site will participate in a training session, review training videos, and submit a set of independently scored examinations that will be used for certification, as was done in the PENUT trial.[2] A re-certification process will be performed at least every 18 months to ensure that primary outcome examiners remain adequately trained throughout the study. Inter-observer variability assessments will be done to determine agreement with gold standard responses. Annotated feedback will be given to examiners regarding items that had a < 85% correct rate. Based on experience in the ELGAN study, we expect agreement rates to exceed 90%.

2.2 Motor Deficit – Gross Motor Functional Classification System (GMFCS)

Gross Motor Functional Classification System (GMFCS) The GMFCS is a well-accepted and validated tool that is used widely to classify motor functional outcomes. Distinctions between levels of motor function are based primarily on functional ability. The GMFCS was used to determine motor outcome in the NINDS-funded Beneficial Effects of Antenatal Magnesium Sulfate (BEAM) trial, as well as in 6 of 7 hypothermia trials for HIE. We will apply the BEAM trial algorithm and definitions when assigning GMFCS scores. Level 0 = normal gait; Level 0.5 =asymmetric gait; Level 1 = walks independently with abnormal gait or requires anklefoot orthosis; Level 2 = cruises, pulls to stand, sits hands free; Level 3 = sits propped on hands only, rolls both ways; Level 4 = sits when supported in lower trunk, has head control, can roll to supine; Level 5 = no head or trunk control, no rolling, little or no voluntary movement. These definitions are modified from the Palisano[3] algorithm adapted by Rosenbaum P and Saigal S for TIPP Trial, and utilized in the NICHD Neonatal Research Network under the direction of Betty Vohr.

2.3 Cognitive Deficit – Bayley III

Bayley III Cognitive Score < 90 Bayley III Cognitive Score[4] is a standard test used to evaluate early cognitive outcomes in high-risk infants. Severity of cognitive deficit is defined by number of standard deviations (SD) below the mean (i.e., severe, moderate, and mild deficit = -3, -2 and -1 SDs below the mean). Since the Bayley III yields higher cognitive scores than the previous version of the Mental Development Index on the Bayley II,[5-8] we have defined cut-offs on the Bayley III that are 5-15 points shifted to the right compared to the Bayley II. We consider any Bayley III Cognitive Score < 90 as abnormal, with mild, moderate, and severe ranges as defined in **Table 2**. These definitions are consistent with findings in a contemporary cohort of infants with HIE who underwent hypothermia,[7] and also consistent with cut-offs used in the PENUT trial. Each site will undergo certification by reviewing a “gold standard” Bayley III assessment performed by our research neuropsychologist. Each site psychologist will

then record a Bayley III assessment of a 2-year-old child and submit the test booklet and video recording for review by the research psychologist for feedback. HEAL CCC leadership has experience in performing multicenter Bayley III certification in other NIH funded studies (e.g., PENUT, TOLSURF), and will use the infrastructure in place for these studies. Only Bayley III examiners who have undergone this training and certification will be allowed to perform the 2-year primary endpoint evaluation.

2.4 Adjudication of Primary Outcome

There may be cases of only a phone follow-up rather than an in-person visit, or otherwise partial data. The CCC will create an Outcomes Adjudication Committee, who will be masked to the treatment assignment of the child. The Committee will receive all available data on the long-term outcome of these toddlers and will assess whether they can definitively assign a primary outcome level and what it is, or they do not have sufficient information to make that determination and therefore the value will be imputed. This model was successfully previously used in the NO CLD clinical trial long-term outcome.

3. Secondary Outcomes

At age 2, we will assess the effect of Epo on secondary outcomes: a) presence of CP, b) severity of motor impairment, c) Bayley III cognitive and language scores, d) epilepsy (i.e., ≥ 2 afebrile, unprovoked seizures), and e) behavioral abnormalities (i.e., attention problems or aggressive behavior) based on the CBCL externalizing score.[9]

Severity of motor impairment will be determined by type of CP and GMFCS level (**Table 1**).

Table 1. Motor outcome - 4 level classification.

	GMFCS						
	0	**0.5**	**1**	**2**	**3**	**4**	**5**
No CP	None	None	Mild	Moderate	Severe	Severe	Severe
HP or DP	Mild	Mild	Moderate	Moderate	Severe	Severe	Severe
QP	Moderate*	Moderate*	Severe*	Severe	Severe	Severe	Severe

** It is unlikely that a child with quadriparetic CP will have a GMFCS of 0-1. However, this scenario is possible in cases of bilateral hemiparesis in which arms are more affected than legs. In such cases, the bilateral nature of the deficit, and the significant neurologic abnormalities that are noted on a standardized neurologic examination, warrant a designation of moderate/severe neurodevelopmental impairment.*

QP: quadriplegic; HP: hemiplegic; DP: diplegic

Severity of cognitive impairment will be determined by Bayley III Cognitive Score (**Table 2**).

Table 2. Cognitive outcome – 4 level classification.

Cognitive Deficit	**Bayley II MDI score (Hypothermia trials)**	**Bayley III Cognitive score (PENUT, HEAL)**
Severe (> 3 SD)	≤ 55	≤ 70
Moderate (2-3 SD)	> 55 and ≤ 70	70-84
Mild (1-2 SD)	> 70 and ≤ 85	85-89
None	> 85	≥ 90

4. Exploratory Outcomes

4.1 Ordinal Exploratory Outcome

To elucidate the effect of Epo on all severities of impairment, we will analyze the effect of Epo on the following 4-level outcome: 1) normal, 2) mild motor and/or cognitive impairment, 3) moderate/severe motor and/or cognitive impairment, and 4) death. Use of this measure allows a detailed assessment of potential shifts in the distribution of outcomes toward improved status associated with treatment.

4.2 Circulating Biomarkers

Inflammation is thought to play an important role in HIE and CP.[10-13] Biomarkers of inflammation that we measure include: interleukin (IL)-1β, IL-6 and IL-8 and TNF-α. Putative biomarkers of brain injury in neonatal HIE include: glial fibrillary acidic protein (GFAP), ubiquitin C-terminal hydrolase-L1 (UCH-L1), S100B, Tau, and neuron-specific enolase (NSE).

4.3 Neuroimaging Outcomes

The HEAL neuroimaging outcome measure is the total MRI injury score based on a previously validated MRI scoring system for HIE.[14] This score indicates the extent of signal abnormality (i.e., 0=none, 1 ≤25%, 2=25%–50%, 3 ≥50%) for each of the deep grey nuclei, PLIC, cerebral cortex, white matter and cerebellum, as well as brainstem (scored on a 0–2 scale) separately for each sequence. 'Signal abnormality' is defined qualitatively as either abnormally low or high signal on T1w and T2w images or areas of restricted diffusion on Trace-weighted or ADC images calculated from the DTI sequence. We do not employ a quantitative threshold for ADC, consistent with the previously validated scoring system.[14,15] Areas with high ADC are not scored as injury for the DTI images, which is designed to measure acute injury; however, correspondent areas of low and/or high signal on the T1w and/or T2w images are scored.

For each ROI, injury is scored qualitatively for each sequence on a 4-point scale based on the extent of signal abnormality (0=none, 1 = <25%, 2=25%–50%, 3 ≥50%), except for the brainstem which is scored on a 3-point scale (0=none; 1=focal; 2=multifocal/widespread).

To compute the MRI injury score, each scan is reviewed independently by two of three experienced readers (AMM, JLW, and RCM), who are blinded to the infant's clinical course, treatment assignment and MRS findings. After the primary review, each scan undergoes final consensus review during which any discrepancies are reviewed by all three reviewers and resolved by consensus. The individual and consensus scores are then classified as none (total=0), mild (1–11), moderate (12–32) or severe (33–138), in accordance with the previously validated MRI injury classification.[14] Inter-rater reliability between the three independent readers will be determined for both the total score and for the categorical classification (none, mild, moderate, severe). For the total score, we will use a general linear mixed effects model to estimate the intraclass correlation coefficient (ICC).[16] For the categorical classification, we will use kappa.[17]

Secondary MRI measures include the MRI injury severity classification discussed above as well as classification of injury pattern and acuity. Injury pattern is classified as: (1) normal MRI (defined as no evidence of injury); (2) central (injury to the BGT±perirolandic cortex); (3) peripheral (injury to the parasagittal cortex and/or WM, i.e, 'borderzone/watershed distribution); (4) global (injury to BGT+ total or near total involvement of cortex±underlying WM); (5) punctate WM Lesions (discrete foci of injury typically 1–10 mm in size localised to the periventricular WM or centrum semiovale); (6) arterial ischaemic stroke (infarct localised to the vascular territory of the middle, anterior or posterior cerebral arteries); (7) other focal lesions (includes venous infarcts, contusions, and unilateral lesions to the BGT,

cortex or white matter not classified elsewhere) and (8) atypical/not otherwise specified (with additional text field for describing the lesion). The scoring system allows for multiple patterns to be coded, which will allow us to determine the frequency of individual patterns as well as the co-occurrence across patterns.

4.4 Hearing and Cortical Visual Impairment

To explore the effect of Epo on sensory deficits that can result from HIE, we will collect 2-year data regarding hearing impairment requiring hearing aids, and if available, information about the presence of cortical visual impairment diagnosed by an ophthalmologist.

5. Data Analyses

5.1 Primary Outcome

The primary analysis will be a test of equality of the rate of the primary outcome (death or neurodevelopmental impairment (NDI)) across the two randomized investigational groups. Specifically, we will use a likelihood ratio test based on logistic regression, with stratification by recruitment center and HIE severity. We will perform modified Intention to Treat (mITT) analysis and expect minimal non-compliance due to the nature of the intervention in relation to in-patient care. For the primary endpoint, we expect uniform and complete ascertainment of death but may not evaluate all subjects for developmental impairment. We plan to perform a primary analysis based on complete cases and will exclude those subjects for whom vital status is known (alive) but NDI cannot be assessed. Sensitivity analysis will use multiple imputation to evaluate the potential impact of any missing data. Secondary analysis will consider an ordered categorical two-year status measure (death/severe or moderate impairment/mild impairment/normal), and analysis will be based on generalized Wilcoxon tests.

5.1.1 Subgroups

Given the a priori hypothesis that treatment effect may differ according to gender or HIE severity we will conduct a pair of subgroup analyses that assesses treatment effects separately for males and for females, and separately for moderate and severe HIE. Subgroup specific treatment effects will be computed and inference will be based on a single covariate-by-treatment test for interaction using logistic regression.

5.2 Secondary Outcomes

Generalized linear regression models will be used for binary and continuous secondary outcomes, adjusted for recruitment site and HIE severity.

5.3 Exploratory Outcomes

Generalized linear regression models will be used for exploratory binary and continuous outcomes, adjusted for recruitment site and HIE severity. Regression models for ordered categorical outcomes (proportional odds models) will be used to provide treatment effect estimates adjusted for recruitment site and HIE severity. Quantitative measures include MRI-based injury score using the Washington University Standardized Scoring System. For these endpoints, differences and 95% confidence intervals in median brain MRI injury scores between groups will be assessed using bootstrapped resampling with replacement.

5.4 Exploratory Prognostic Biomarkers

We will consider two main classes of potential predictors of 24-month status: neuroimaging measures and inflammatory markers. Interest is in the prognostic potential of individual and/or combined biomarker measurements. Given that the primary outcome is a binary measure (NDI), we will evaluate the predictive potential of individual quantitative measures using ROC curves showing the full potential of sensitivity and specificity across marker cut points. We will compute ROC curves for the (4) primary neuroimaging measures, and separately for individual inflammatory markers. We will derive two multivariate predictive models: using the inflammatory markers and using the MRI and MRS measures. We will use AIC and 10-fold cross-validation to develop and validate predictive models. A final multivariate model will combine markers from both MR and inflammatory measures, and 10-fold cross-validation will permit inference in the incremental value of adding markers in combination by comparing ROC curves and associated area under the ROC curve (AUC). Evaluation of whether treatment modifies the prognostic potential of biomarkers can be conducted by testing for the interaction between treatment status and individual biomarkers in predictive models for 2-year outcomes.

5.5 Safety Data

Clinical safety data includes serious adverse events (SAEs) and clinical laboratory markers both from the hospitalization and study intervention period and from the long-term follow-up period. Because these are critically ill newborns, we anticipate that most safety events will occur during the initial inpatient and treatment period. Safety event rates will be tabulated by study group and compared separately for each SAE using a multivariable logistic regression model, with adjustment for randomized treatment group, and randomization stratification factors of clinical recruitment site and HIE severity (moderate/severe). We will additionally create a per-patient aggregate count of SAEs and evaluate the rates between groups using a Poisson regression model with robust standard errors and adjustment for randomization stratification factors. For any safety events that occur infrequently (<2%), we will use Fisher's exact test to compare the rates between treatment groups. All safety events identified post-birth will be included in the primary analysis, while the secondary analyses will evaluate safety events occurring post-treatment with the first study drug dose as well as those events assessed as being possibly related to the study drug. Statistical significance is defined conservatively at the alpha=0.05 level with no correction for multiple outcomes.

5.6 Laboratory Data

Laboratory tests of organ injury are measured at baseline and at age 2-3 days as part of routine care. We will compare laboratory test data measured between study day 2 and 3 between treatment groups using analysis of covariance (ANCOVA) regression model, with adjustment for treatment group, the laboratory value measured at baseline, and the time between laboratory measurements. We will use a similar analytic approach to compare vital signs and growth parameters between treatment groups.

5.7 Circulating Biomarkers

Inflammation is thought to play an important role in HIE and CP.[10-13] Biomarkers of inflammation that we measure include: interleukin (IL)-1β, IL-6 and IL-8 and TNF-α. Putative biomarkers of brain injury in neonatal HIE include: glial fibrillary acidic protein (GFAP), ubiquitin C-terminal hydrolase-L1 (UCH-L1), S100B, Tau, and neuron-specific enolase (NSE). We will select a random subset of 200 subjects (100 treated and 100 controls with each group, for example, including both moderate and severe HIE) to measure circulating biomarkers of inflammation and brain injury. We will collect 3 plasma samples from each infant at the following time points based on hour of age: < 24 hours; Study Day 2; Study Day 4. Our analysis will focus on time-specific comparisons of the mean biomarker measure across treatment

groups using appropriate regression methods while controlling for site and HIE severity. In addition, we will conduct longitudinal analysis using linear mixed models[18] that permit an omnibus test across all four measurement times and allow inference on differential rates of change across treatment groups.

5.8 Neuroimaging Outcomes

The primary neuroimaging outcome measures are: (1) the total injury score (derived from the HEAL/Wash U MRI scoring system above) and (2) the ratio of lactate/NAA determined for the left thalamus and parietal white matter from the quantitative MRS data. For the MRI injury score and the lactate/NAA ratio, we will use linear regression to compare Epo treated patients to controls while adjusting for site and HIE severity since these are factors used to stratify the randomization. If we have missing data due to either early patient death or other factors (e.g., family declines MRI or infant is unable to complete all the sequences), we will estimate the resultant sampling bias by comparing our final neuroimaging sample to the overall HEAL sample with regard to patient demographics and primary outcome data.

6. Compliance, Retention, and Missing Data

6.1 Adherence and Retention

Our major analyses are based on the modified ITT principle. We do not anticipate that non-adherence will be a major issue since the treatment is directly observed during a short in-hospital time frame. We will assess non-compliance, with particular focus on study treatment dosing and timing. If there are more than minimal issues, that will justify quantifying and characterizing non-adherence and doing a per-protocol analysis. In addition, site selection included having a committed neonatology follow-up program, and we expect >90% retention.

6.2 Missing Data and Dropouts

We will strive to sustain excellent participant involvement throughout the study and we have achieved 90%+ follow-up rates in numerous prior studies. The UW DCC will generate automated nightly reports, available to staff at the study sites, identifying these fields with a request to discuss and prevent further missingness. Important data elements will be prospectively monitored to examine patterns of missingness.

6.3 Handling

In final manuscripts and analyses, the number of non-responders will be enumerated by study arm according to CONSORT guidelines. We will conduct a missing data analysis to describe and characterize enrolled participants who do not provide further response due to attrition or dropout. We will also conduct sensitivity analyses using 10-fold multiple imputation to assess the robustness of the results when missing data are imputed.

6.4 Statistical Uncertainty

Given the type of missing data we expect in the proposed study, missing completely at random (MCAR) or missing at random (MAR), both methods we propose to utilize for missing data properly account for statistical uncertainty due to missingness and will provide accurate confidence interval coverage.

6.5 Sensitivity Analysis

We will assess the sensitivity of inferences made from missing data methods first by using the two previously described methods for dealing with missing data, and secondly by imputing missing data under both pessimistic and optimistic scenarios to provide bounds on the statistical uncertainty. The

characteristics of non-responders will be summarized in our final report where we will present the sensitivity of the treatment effect due to missing data.

7. Sample Size and Accrual

7.1 Primary Outcome Sample Size Calculations

Our proposed HEAL sites report an overall mortality rate of 14%. Using three large sites that participated in the phase II study (UCSF, Wash U, CNMC) we can also estimate the rates of neurodevelopmental impairment: death = 14%; moderate-severe impairment = 18%; mild impairment = 17%; and normal = 51%. Therefore, we anticipate a control primary outcome rate of 49% (death or NDI).

7.1.1 Non-human Data Informing Treatment Effect Size

A recent study with nonhuman primates (Macaca nemestrina) compared animals experiencing 15-18 minutes of umbilical cord occlusion that were then treated with either saline (n=14), therapeutic hypothermia (n=9), or therapeutic hypothermia and multiple doses of Epo.[19] Among animals treated with saline 8/14 = 57% died or had NDI. Among animals treated with hypothermia (HT) alone 7/9 = 78% died or had NDI, while among animals treated with hypothermia and Epo (HT+Epo) only 5/12 = 42% were observed to die or have NDI. Results from Figure 2 of Traudt et al.[20] show the number of animals in each outcome category by treatment group. These data suggest a risk ratio of 0.53 (95% CI: 0.25, 1.14) comparing HT+Epo to HT alone, and a risk ratio of 0.73 (95% CI: 0.32, 1.64) comparing HT+Epo versus saline. Therefore, animal data support an Epo effect that optimistically corresponds to a 50% reduction, and that is conservatively associated with a 27% reduction in the primary outcome rate.

7.1.2 Phase I Data Informing Treatment Outcome Rates

Given that the primary outcome is based on a 22-26 month assessment (and for extenuating circumstances such as COVID-19 restrictions, up to 36 months of age) we can rely on a recently completed long-term follow-up from a phase I study[21] in which 24 cooled infants were given multiple doses of Epo ranging from 250U/kg to 2500U/kg24. At 22-26 months of age n=22 subjects were followed, and 0/22 subjects died, 1/22 had moderate-severe NDI, and 6/22 had mild NDI. This study suggests an overall primary outcome rate of 7/22 = 31.8% (exact confidence interval = 14%-55%) for our planned intervention group. We recognize that the dosing of Epo was not optimized in the phase I trial, and our proposed study will use multiple doses of 1000U/kg for all subjects which has been shown to yield plasma concentrations observed to be neuroprotective in animal studies.

Table 3. Distribution of outcomes in previous studies

	NEATO Epo (n=22)	NEATO Placebo (n=25)	Bayley III Cognitive	Trivedi et al. (submitted)
Injury Severity			**Mean**	**Std. Dev.**
None (0)	8 (36%)	3 (12%)	96	(8)
Mild (1-11)	13 (59%)	11(44%)	100	(13)
Moderate (12-32)	1 (5%)	6 (24%)	85	(18)
Severe (>32)	0 (0%)	5 (20%)	76	(19)
Predicted Bayley III (Mean (SD))	97.9 (11.7)	91.1(15.2)		
Predicted %<90	25.1%	47.1%		
Predicted death/NDI	34.1%	54.5%		

7.1.3 Phase II Data Informing Treatment Outcome Rates

In our phase II NEATO trial, a total of n=50 subjects were randomized to Epo (n=24) or placebo (n=26). We find substantial differences in the MRI injury severity distribution with 95% of Epo treated subjects having no injury or mild injury as compared to only 56% of control subjects. Using a recently submitted study from Trivedi, et al., we can then link the MRI injury severity category to expected Bayley III Cognitive scores at 24 months. Using our NEATO data and Trivedi's data we calculate a predicted mean (SD) of 97.9 (11.7) among Epo treated subjects, and a mean (SD) of 91.1 (15.2) among controls. The predicted 6.8-point mean difference for Bayley III Cognitive scores is consistent with our observed 6.3-point difference in mean WIDEA scores at 6 months among NEATO subjects. Predicted Bayley III Cognitive distributions lead to an expected 25.1% of subjects with a cognitive score of < 90 among Epo treated subjects, and 47.1% among controls. Incorporating expected death rates of 12% and 14% respectively for Epo treated and controls leads to an expected primary outcome rate of 34.1% among treated and 54.5% among controls (**Table 3**). We acknowledge that our primary outcome is based on both Bayley III Cognitive scores, and clinical assessments of CP and GMFCS level, but expect Bayley status to be the major case indicator.

In addition, Cheong et al.[22] evaluated the correlation between MRI findings obtained within 10 days of birth and 2-year clinical status outcomes for participants in the ICE trial (Infant Cooling Evaluation). Specifically, basal ganglia and thalamus (BGT) injuries were classified as abnormal if moderate/severe abnormalities were noted on T1- and T2-weighted images. In our phase II trial, we find only 4.5% (1/22) among Epo treated subjects have BGT injury as compared to 20% (5/25) among control subjects. Cheong et al. estimates that the probability of death/NDI at two years is 88% (PPV) for BGT abnormal subjects as compared to 32% (1-NPV) for BGT normal subjects. Applying these rates to our phase II BGT results yields **expected death/NDI rates of 34.5% among Epo treated**, and 43.2% among controls. For the ICE trial, the overall death/NDI rate for ICE hypothermia subjects was 51%, which is approximately the same as our expected control rate. Therefore, based on animal data, phase I data, and projections from phase II data we expect primary outcome rates from 31-35% among Epo treated subjects with a protective relative risk of 0.65 to 0.71.

7.1.4 Power and Sample Size for Primary Outcome

In order to determine the necessary sample size for efficacy evaluation we need to formulate assumptions for the primary outcome rate in the Epo treated and control groups. The primary outcome measure is the composite rate of death or NDI, and current cohort studies suggest that the primary outcome occurs among 49% of infants treated with hypothermia alone (standard of care). Based on human data presented in Rogers et al.[23] (2014), animal studies including Traudt et al.[20], and the NEATO phase II data we assume that 33% of treated infants will die or have NDI, corresponding to a relative risk of 0.67. Assuming an intervention rate of 33% yields greater than 90% power, while we have 88% power for an alternative of 34% (**Table 4**). In order to compute power, we assume a 90% follow-up rate with n=225/250 subjects evaluated in each arm.

Table 4. Power analysis for the primary outcome assuming n=500 patients randomized and 10% loss to follow-up

Control	Intervention	Relative Risk	Power
49%	32%	0.65	95%
49%	33%	0.67	92%

49%	34%	0.69	88%
49%	35%	0.71	83%

7.2 Sample size for efficacy secondary outcomes

Our key secondary long-term outcome is an ordered categorical 24-month status measure that classifies subjects as: dead; moderate or severe impairment; mild impairment; and normal. Use of this measure allows a detailed assessment of potential shifts in the distribution of outcomes toward improved status associated with treatment. Statistical analysis of this outcome will use a generalization of the Wilcoxon test that controls for recruitment site and HIE severity. Regression models for ordered categorical outcomes can also be used to provide adjusted treatment effect estimates.

To consider power we have used 2013 data from three large sites participating in the phase II trial (UCSF, Wash U, CNMC) to estimate the distribution of outcomes within the control group. We calculate power for the stratified Wilcoxon test assuming a set of alternatives that are consistent with the assumption of a 33% rate of death or impairment (Scenarios 1 and 2, **Table 5**) or a 32% rate (Scenarios 3 and 4, **Table 5**) associated with intervention that was used to power the primary analysis. We assume a small effect on the death rate, and compute power for alternative shifts in the distribution of outcomes. For example, in Scenario 1 we assume no reduction in the death rate, and an 8% reduction in the moderate/severe category (relative risk = 0.55), and an 8% reduction in the rate of mild impairment (relative risk = 0.50) which yields 81% power using a two-sided alpha=0.05 test. Scenario 3 considers a larger reduction in the rate of moderate/severe and a smaller reduction in the mild category and yields 90% power. Scenarios 3 and 4 consider a 32% overall rate of impairment or death and yield 90% or greater power. Therefore, our study is adequately powered to detect modest but clinically important shifts in the outcome distribution.

Table 5. Power analysis for a four-level outcome measure with n=500 patients randomized and assuming 10% loss to follow-up

	Neurodevelopmental Impairment				
	Normal	**Mild**	**Moderate/Severe**	**Mortality**	**Power**
Placebo Arm	51%	17%	18%	14%	--
Epo Arm					
Scenario 1	67%	9%	10%	14%	81%
Scenario 2	67%	11%	10%	12%	86%
Scenario 3	68%	9%	9%	14%	86%
Scenario 4	68%	11%	9%	12%	92%

Given the a priori hypothesis that treatment effect may differ according to gender or HIE severity we will conduct a pair of subgroup analyses that assesses treatment effects separately for males and for females, and separately for moderate and severe HIE. Subgroup specific treatment effects will be computed and inference will be based on a single Covariate-by-Treatment test for interaction using logistic regression.

7.3 Sample size for MRI/MRS biomarkers

For imaging measures, we will have data for all subjects. Assuming a 10% missing data rate we have 80% power to detect a difference in the mean of secondary outcomes across treatment groups of 0.26 SDs. For the inflammation and brain injury biomarkers, we will have a total of 200 subjects and have 80% power to detect a mean difference across the treatment groups of 0.4 SDs.

7.4 Sample size for plasma biomarkers of brain injury and inflammation

A sample size of 91 neonates per group would have been required to detect a 0.5 SD difference on the log scale with a single observation per neonate with 80% power while controlling for a Type I error of 0.0125 (4 markers, Bonferroni). However, since we have 4 samples per neonate, a sample size of 100 neonates per treatment group (with correlated 400 samples per group) will be more than sufficient to attain >80% power.

7.5 Achieving the sample size

The plan is to enroll and randomize 500 patients over a 36-month period. The 17 sites that were included in the original study proposal cooled a total of 530 HIE infants in the year 2014. From the NEATO phase II randomized controlled trial study, we estimate that at least 40-45% of these infants would be eligible and will consent for a phase III trial, yielding at least 16 enrolled infants per month. With a further conservative estimate of 14 per month allowing for slower enrollment during start-up, we can enroll 500 patients in 36 months.

8. Data Monitoring

We will monitor the accuracy of data entry by the sites both internally and externally. We will review study data on arrival for completeness. We will then subject each submitted data set to a set of preliminary checks to search for values that are out-of-range or otherwise inappropriate. Using the Patient Monitoring Report, a subset of all data points in the CRFs will be compared with the medical record for 20-25% of enrolled subjects. Any outstanding data queries will be resolved with the research coordinator at the time of the site-monitoring visit. After each study site monitoring visit, a report will be prepared and copies sent to the Study File, the study PIs (Y. Wu and S. Juul), the site PI, and the site coordinator. The quality and completeness of other deliverables (blood samples, MRIs) will be monitored.

9. Data and Safety Monitoring Plan (DSMP)

An NINDS-appointed independent Data and Safety Monitoring Board (DSMB) will review the accruing data to: 1) ensure that the study is adequately enrolling; 2) to ensure that there are no serious safety concerns; and 3) to assess whether the study efficacy appears overwhelming. The DSMB will be assigned by NINDS. The research coordinator at each site will monitor each subject weekly for the presence of any complications. Serious adverse events will be brought to the attention of the DSMB, and if appropriate, the IRB, in writing. An independent medical monitor will review all cases of serious adverse events.

As part of this DSMP, we will perform continuous and interim analysis of accruing safety data. We have defined potentially treatment (Epo) related serious adverse events (SAEs) that will be monitored throughout the course of the study. Specifically, for SAEs we will compare absolute rates to expected rates based on published data for similar newborns and will seek careful DSMB review and guidance when observed rates exceed pre-specified thresholds. The following SAEs will be prospectively reviewed by the DSMB:

- Systemic hypertension requiring treatment
- Polycythemia (hematocrit > 65%)
- Major venous or arterial thrombosis (clot)
- Disseminated Intravascular Coagulation (DIC)
- Pulmonary Hypertension
- Intracranial Hemorrhage

- Cardiopulmonary Arrest
- Death
- Other unexpected life-threatening event (unexpected for HIE or Epo drug profile)

In addition, at planned interim analysis we will formally compare the event rates across the two treatment groups using appropriate small sample methods such as Fisher's exact test. The DCC PI will remain masked to assignment while the study staff statistician will not.

The primary outcome of the study is a composite endpoint of mortality or NDI at 24 months of age. Therefore, monitoring the primary outcome for treatment efficacy or futility is challenging. Based on enrollment plans, most patients will have been randomized by the time NDI is assessed at 24 months for any participants, and therefore we do not expect to be able to conduct a first interim analysis prior to the completion of enrollment. We expect to have the primary outcome evaluated on the first quarter of subjects (n=125) after 34 months of recruitment, at which point we expect to have randomized n=450 of the total n=500 subjects (90% enrollment completed). Therefore, any actions that a DSMB might take to prevent subsequent patients from receiving an ineffective treatment (futility) or to make available a useful treatment (efficacy) will not have a direct impact on patients participating in this study. As a surrogate for long-term 24-month clinical efficacy on NDI, one alternative would be to monitor directly early MRI results as intermediate outcomes or surrogates. However, a large study published by Cheong et al.,[22] found that early MRI measurements had poor sensitivity (27-60%) for accurately predicting death or NDI at 2 years. Therefore, we do not recommend using early MRI measures as a surrogate for the long-term outcome for monitoring treatment efficacy. Given that we are not conducting interim analysis directly on the primary outcome measure for efficacy or futility, we do not adjust sample size or statistical power to account for interim alpha spending.

Our primary objective for interim analyses is therefore to allow for careful and continued monitoring of mortality and safety outcomes. We propose to conduct formal statistical analysis and inference for mortality at three interim and one final analysis time. We will continue to present mortality and SAE data from all available follow-up data, but we expect most treatment-related safety events to occur within the first three months of follow-up. We will conduct formal safety evaluation at 6, 12, 18, and 24 months following the start of enrollment, where approximately 25, 50, 75, and 100% of the study cohort will have been randomized and followed for at least 3 months. As part of each interim analysis, we plan to monitor mortality as a primary safety endpoint and will control the overall significance level using O'Brien-Fleming boundaries (net alpha=0.05 significance, accounting for three interim and one final analysis). The DSMB will also monitor all other SAEs utilizing the same O'Brien-Fleming sequential monitoring boundaries (**Table 6**) without further adjustment for multiple comparisons but allowing for flexibility to continue the study if the O'Brien-Fleming boundary is reached on a secondary SAE endpoint.

Table 6. O'Brien-Fleming Monitoring Boundaries

Enrollment	Monitoring p-value: Death
25%	0.000014734
50%	0.0030359
75%	0.016248
100%	0.030701

Data Safety Monitoring Schedule: Our target enrollment is n=500 which is expected to accrue during the first 4.0 years of the trial. Therefore, we will enroll approximately 83 subjects every 6 months. Our planned DSMB safety analyses will occur every 6 months after trial initiation. At the first safety evaluation, we expect to have 83 subjects enrolled, but would only have discharge outcomes on the first 60 subjects (**Table 7**). Each subsequent 6-month period would increase the number of babies by 83 leading to the following cumulative

number of subjects available for analysis. Safety evaluation will be based on all available follow-up, but we expect most SAEs to occur during the neonatal hospitalization and therefore within one month of age (2-4 weeks) in most cases.

Table 7. Cumulative number of subjects evaluated for safety events and for the long-term efficacy outcome.

Month	6	12	18	24	30	36	42	48	54	60	66
Safety (1 mo)	60	143	226	309	392	475	500				
Efficacy (24 mo)	0	0	0	0	60	143	226	309	392	475	500

10. References

1. Kuban KC, 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.
2. Kuban KC, O'Shea M, Allred E, et al. Video and CD-ROM as a training tool for performing neurologic examinations of 1-year-old children in a multicenter epidemiologic study. *J Child Neurol.* 2005;20(10):829-831.
3. 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.
4. Ohls RK, Kamath-Rayne BD, Christensen RD, et al. Cognitive outcomes of preterm infants randomized to darbepoetin, erythropoietin, or placebo. *Pediatrics.* 2014;133(6):1023-1030.
5. Anderson PJ, De Luca CR, Hutchinson E, Roberts G, Doyle LW, Victorian Infant Collaborative G. Underestimation of developmental delay by the new Bayley-III Scale. *Arch Pediatr Adolesc Med.* 2010;164(4):352-356.
6. Vohr BR, Stephens BE, Higgins RD, et al. Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *J Pediatr.* 2012;161(2):222-228 e223.
7. Moore T, Johnson S, Haider S, Hennessy E, Marlow N. Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *J Pediatr.* 2012;160(4):553-558.
8. Chalak LF, DuPont TL, Sanchez PJ, et al. Neurodevelopmental outcomes after hypothermia therapy in the era of Bayley-III. *J Perinatol.* 2014;34(8):629-633.
9. Achenbach TM, Rescorla LA. Manual for the ASEBA Preschool Forms & Profiles. In. *Child Behavior Checklist for Ages 1.5-5*. Burlington, VT: Research Center for Children, Youth & Families; 2000.
10. Wu YW, Escobar GJ, Grether JK, Croen LA, Greene JD, Newman TB. Chorioamnionitis and cerebral palsy in term and near-term infants. *JAMA.* 2003;290(20):2677-2684.
11. Wu YW, Croen LA, Torres AR, Van De Water J, Grether JK, Hsu NN. Interleukin-6 genotype and risk for cerebral palsy in term and near-term infants. *Ann Neurol.* 2009;66(5):663-670.
12. Yang L, Sameshima H, Ikeda T, Ikenoue T. Lipopolysaccharide administration enhances hypoxic-ischemic brain damage in newborn rats. *J Obstet Gynaecol Res.* 2004;30(2):142-147.
13. Badawi N, Kurinczuk JJ, Keogh JM, et al. Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study [see comments]. *BMJ.* 1998;317(7172):1554-1558.
14. 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.
15. Wu YW, Mathur AM, Chang T, et al. High-Dose Erythropoietin and Hypothermia for Hypoxic-Ischemic Encephalopathy: A Phase II Trial. *Pediatrics.* 2016;137(6).

16. Nakagawa S, Johnson PCD, Schielzeth H. The coefficient of determination R(2) and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface.* 2017;14(134).
17. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med.* 1994;13(23-24):2465-2476.
18. Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Second Edition ed: Oxford University Press; 2002.
19. Traudt CM, McPherson RJ, Bauer LA, et al. Concurrent erythropoietin and hypothermia treatment improve outcomes in a term nonhuman primate model of perinatal asphyxia. *Dev Neurosci.* 2013;35(6):491-503.
20. Traudt CM, Juul SE. Erythropoietin as a neuroprotectant for neonatal brain injury: animal models. *Methods Mol Biol.* 2013;982:113-126.
21. Wu YW, Bauer LA, Ballard RA, et al. Erythropoietin for neuroprotection in neonatal encephalopathy: safety and pharmacokinetics. *Pediatrics.* 2012;130(4):683-691.
22. Cheong JL, Coleman L, Hunt RW, et al. Prognostic utility of magnetic resonance imaging in neonatal hypoxic-ischemic encephalopathy: substudy of a randomized trial. *Arch Pediatr Adolesc Med.* 2012;166(7):634-640.
23. Rogers EE, Bonifacio SL, Glass HC, et al. Erythropoietin and Hypothermia for Hypoxic-Ischemic Encephalopathy. *Pediatr Neurol.* 2014:In Press.