RTI INTERNATIONAL | *turning knowledge into practice*

A Multi-center Randomized Trial of Laparotomy vs. Drainage as the Initial Surgical Therapy for ELBW Infants with Necrotizing Enterocolitis (NEC) or Isolated Intestinal Perforation (IP): Outcomes at 18-22 months Adjusted Age.

Short Title: Necrotizing Enterocolitis Surgery Trial (NEST)

Statistical Analysis Plan

Protocol Date: September 8, 2009

SAP Author: Barry Eggleston

SAP Version: Version 3

Original SAP Date: August 1, 2013

Date of last revisions: August 23, 2019

Abbreviations

ELBW	Extremely Low Birth Weight
IP	Isolated Intestinal Perforation
NDI	Neurodevelopmental impairment
NEC	Necrotizing Enterocolitis
NEST	Necrotizing Enterocolitis Surgery Trial
NICHD	*Eunice Kennedy Shriver* National Institute of Child Health and Human Development
NRN	Neonatal Research Network
SAP	Statistical Analysis Plan

Preface

This Statistical Analysis Plan (SAP) describes the planned analysis and reporting for the Necrotizing Enterocolitis Surgery Trial (NEST) completed among NRN sites and sponsored by NICHD. This randomized clinical trial is being completed to assess the safety and efficacy of initial laparotomy vs. initial drainage for the treatment of NEC or IP among ELBW infants.

The following documents were reviewed in preparation of this SAP:

- NEST protocol, dated September 8, 2009
- NEST operations manual
- NEST Case report forms (CRFs)

The reader of this SAP is encouraged to also read the NEST protocol and operations manual for details on the conduct of this study, as well as, the timing and operational aspects of clinical assessments for an infant enrolled in this study.

Purpose and Scope of SAP

The purpose of this SAP is to outline in more complete detail the planned analyses as specified in the NEST protocol. The planned analyses identified in this SAP will be included in future manuscripts as deemed necessary by the NEST subcommittee. It is not the intent of this SAP to identify all possible analyses that may be completed using the NEST database. The intent of this SAP is only to describe the analyses required to address the objectives/specific aims/hypotheses explicitly stated in the protocol. Additional exploratory analyses not identified in the protocol or not included in this SAP may be performed as desired by the NEST subcommittee. Any additional analyses not identified in this SAP can be documented per the instructions of RTI's NRN principal investigator.

Study Directives and Endpoints

Study Objectives, Specific Aims, and Hypotheses

Primary Objective

The primary objective is to assess, as rigorously as possible, whether the initial surgical treatment of ELBW infants who have NEC or IP and require surgical treatment should be a laparotomy or percutaneous drain.

Secondary Objectives

1. Determine whether or not the more desirable surgical procedure depends on diagnosis, by assessing whether outcomes at 18-22 months are affected by an interaction between surgical treatment (laparotomy and drainage) and diagnosis (NEC or IP).
2. Determine the accuracy of the surgeon's preoperative diagnosis (NEC or IP) in patients that undergo a laparotomy.
3. Assist in designing future Network trials, to gain experience with three strategies that have been proposed for use when it is difficult or impossible to conduct conventionally powered randomized trials, by completing secondary analyses using Bayesian methods, propensity scoring, and methods relevant to comprehensive cohort designs.

These objectives have been expanded into the following aims:

Primary Specific Aim

1. Using conventional frequentist analyses assess the relative risk for survival without NDI at 18-22 months corrected age, with initial laparotomy relative to initial drainage among ELBW infants who undergo surgical treatment of NEC or IP.

Secondary Specific Aims

1. Determine whether analyses using propensity scoring and using conventional risk adjusted relative risk support the relative risk among randomized infants and the generalizability of the results to nonrandomized infants.[1]
2. Compare initial laparotomy to initial drainage in order to assess the relative risk for secondary outcomes assessed during the neonatal period (e.g., death; death or prolonged parenteral nutrition; specific surgical complications) and at follow-up (e.g. each component of NDI; Bayley III scores).
3. Compare initial laparotomy to initial drainage in order to assess the relative risk for surgical complications (e.g. procedure-related liver hemorrhage; wound dehiscence; intestinal stricture, requiring operation; intestinal fistula);
4. Assess whether there is evidence that the preferred surgical treatment should differ by preoperative diagnosis (NEC or IP) by evaluating whether there is a statistical interaction between surgical treatment and preoperative diagnosis.
5. Develop improved methods to distinguish ELBW infants with NEC from those with IP.
6. Develop improved methods to define the prognosis of individual infants presenting with signs of NEC or IP.
7. Provide information that would be deemed useful by the subcommittee for any other simultaneous clinical trials.

Finally, the protocol extracts the following hypotheses from the above objectives and specific aims.

Primary Hypothesis

1. Initial laparotomy rather than drainage will result in a higher rate of survival without NDI at 18-22 months adjusted age among ELBW infants presented with signs of NEC or IP.[2]

[1] The risk stratification formula was primarily developed in a prospective observational study of NEC previously conducted by the NRN and includes the following variables: birth-weight, gestational age, vasopressor requirement (yes/no), if infant is on high frequency (oscillating or jet) ventilation or not (any modality of conventional ventilation or no mechanical ventilation), pH, Fi02, and preoperative diagnosis (NEC or IP).

Secondary Hypotheses

1. The generalizability of the results in randomized patients to nonrandomized patients will be supported by analyses using conventional risk adjusted relative risk among nonrandomized patients (as in a comprehensive cohort design) and by analyses using propensity scoring.
2. The laparotomy group will have a similar or better outcome than the drainage group with respect to other outcomes during the neonatal period (death; death or prolonged parenteral nutrition; death or prolonged hospital stay) and at follow-up (e.g. each component of NDI; anthropometry).
3. The proportion of infants who have surgical complications (e.g. procedure-related liver injury, wound dehiscence, intestinal stricture or fistula) will be similar in the two treatment groups.
4. There will be no significant statistical interaction between treatment and diagnosis.[3]
5. The trial data will provide improved methods to distinguish ELBW infants with NEC from those with IP.
6. The trial data will provide improved methods to assess the prognosis of ELBW infants presenting with signs of NEC or with IP.

No hypotheses related to Bayesian analyses were specified in the protocol, but Bayesian analyses will be utilized in a secondary analysis of the primary hypothesis and in other secondary analyses related to the above secondary hypotheses as specified in this SAP. Given the non-specificity of secondary aims 5 through 7 as well as secondary hypotheses 5 and 6, this SAP will not address analysis methods related to distinguishing ELBW infants presenting with NEC from those with IP, prognosis of ELBW infants or identification of information useful in other trials. The analysis methods for these exploratory analyses will be documented and reported as instructed by the coordinating center PI. Any results from these unplanned analyses will also be clearly identified in any resulting peer reviewed manuscript as a post-hoc/exploratory analysis.

---

[2] The trial was powered to test this hypothesis only.

[3] Because the trial was not powered to test for an interaction, "significant statistical interaction" will not be interpreted in this SAP to imply the highly technical statistical concept of "statistical significance" but interpreted to imply the more informal concept of "clinical significance."

Study Endpoints (Target Variables)

Primary Target Variable

The primary outcome variable is death or neurodevelopmental impairment (NDI) at 18-22 months, **see variable Patient.DeathOrNDI in DataDictionary.docx**. NDI is defined as meeting any of the following criteria: moderate to severe cerebral palsy (CP) with Gross Motor Function Classification System (GMFCS) Level >= 2, a Bayley III Cognitive score < 85, < 20-200 bilateral vision, and/or permanent hearing loss that does not permit the child to understand the directions of the examiner and communicate despite amplification

Death will be identified by looking into the NEST status form dataset, NEST SAE form, GDB (both regular and special NEST version for outborn babies), and the neonatal follow-up database, **see variable Patient.Any_Death in DataDictionary.docx.**

The components of NDI are defined as follows:

Cerebral Palsy: Definite abnormalities in classical neuromotor exam, including tone, deep tendon reflexes, coordination and movement, coupled with a delay in motor milestones with a disorder of motor function. Severity of CP is classified according to the GMFCS level. Moderate CP is defined as GMFCS level 2 or 3; severe CP is defined as GMFCS level 4 or 5. **The indicator variable for moderate or sever CP is Patient.CP_MODSEV. The indicator variable for Gross Motor Function Classification System (GMFCS) Level >= 2 is Patient.GMFCS_GTE2. Two other variables for CP are Patient.CP_GRADE and Patient.Any_CP.**

Psychometric testing: Cognitive outcome at 18-22 months will be assessed by certified psychologists or psychometricians using the BSID III. For this study, the cognitive subtest (mean±SD: 100±15) will be administered. Scores on the cognitive subtest of <70 (>2 SD below the mean) is consistent with severe cognitive impairment, but in this study a cutoff of 85 will be used.[4] **The indicator variable for a low Bayley II score is Patient.LowBayley.**

[4] An important issue (for this trial and for neurodevelopmental assessment of premature infants in general) is the potential difference in sensitivities of the Bayley II and III products in defining NDI. Some authors have reported recently that the Bayley III likely underestimates NDI, compared to the Bayley II. Because of these issues, for this trial a Bayley III score of <85 will be one of the conditions defining NDI (in addition to those above).

Sensory deficits are identified through specific question concerning vision and hearing. Children with no useful vision in either eye is consistent with a refraction definition of <20/200 (legally blind). **The indicator variable for blindness is Patient.BILATERAL_BLINDNESS.** Permanent hearing impairment that does not permit the child to understand the directions of the examiner and communicate despite amplification is consistent with severe auditory impairment and would be adequate to meet the definition of NDI. **The indicator variable for deafness is Patient.DEAF.**

NDI Adjudication

For cases where we don't have enough information from the follow up visit to assign a level of NDI, adjudication will be performed. In most cases this happens when the Bayley score is missing and the other components are negative. For such situations, we will convened a subcommittee, and asked them to review the neurologic assessment and follow up visit assessments for each of these babies in a blinded fashion (no treatment group) to determine if they can discern, based on all the available information, whether:
(a) NDI is likely present,
(b) NDI is likely absent, or
(c) NDI cannot be determined.
Once adjudication is complete, we will conduct the primary and affected secondary analyses incorporating this information (treating the last category as missing data). In addition to this analysis, sensitivity analyses will be completed to determine whether the adjudicated outcomes change any results.

Appropriateness of Primary Target Variable

Until recently, all studies evaluating laparotomy and drainage have focused on death and other outcomes occurring in the immediate post operative period. An editorial in the New England Journal of Medicine pointed out the limitations of current studies evaluating surgical therapies in NEC and highlighted the importance of measuring neurodevelopmental outcome beyond nursery discharge. The previously completed observational study by the Network is the only prospective study addressing neurodevelopmental outcome in addition to death at a long-term time point such as 18-22 months.

Results from the Network's observational study suggested that laparotomy may improve long term outcome as measured by death or NDI (risk adjusted odds ratio for death or disability (0.55; 95% CI= 0.18 – 1.67), so the NEST subcommittee considered it very important to study death and NDI at a time point beyond discharge.

Secondary Target Variables

Secondary outcomes variables will include surgical complications,

- such as wound dehiscence, (**variable name is Patient.wound_dehis**)
- intestinal stricture or fistula, (**variable name is Patient.intest_strict and Patient.Fistula)**
- procedure-related liver hemorrhage, (**variable name is Patient. LiverHemorrhage**)
- number of surgical procedures, (**variable name is Patient.TotalSurgeries**)
- sepsis episodes, (**variable name is Patient.GDB_SEPSIS**)
- duration of parenteral nutrition, (**variable name is Patient.GDB_PARDAYS**)
- parenteral nutrition associated cholestasis, (**variable name is Patient.PN_assoc_chol**)
- length of hospital stay, (**variable name is Patient.GDB_HOSPSTAY**)
- rehospitalizations, (**variable name is Patient.?**)
- each component of the primary outcome;
    - Cerebral Palsy classification (GMFCS Level >= 2) [**classification variables are Patient.GMFCS_GTE2 and Patient.Any_CP, while the raw variable for GMFCS is Patient.EFGROSS**.]
    - BSID III score (Bayley III Cognitive score < 85) [**classification variable is Patient.LowBayley, while the raw variable is Patient.NF9ABSCC**.]
    - Vision assessment (< 20-200 vision or legally blind) [**classification variable is Patient.BILATERAL_BLINDNESS, while the raw variables are Patient.EFVISNL and Patient.EFVISNR**.]
    - Hearing assessment (permanent hearing loss) [**classification variable is Patient.DEAF, while the raw variable is Patient. EFHEARIM**.]

Study Methods

Overall Study Design and Plan

Similar to a Comprehensive Cohort Design, the study design is an unmasked randomized trial accompanied by a nonrandomized cohort. Data from the nonrandomized cohort will be analyzed and results compared to similar results derived from analyses of randomized cohort data. Covariate adjustment methods, and propensity scores will be utilized in an attempt to adjust for selection bias in the nonrandomized cohort. The nonrandomized cohort is not a representative sample of all eligible infants not randomized, but it is a sample of eligible infants whose surgeon, neonatologist, or parents refused randomization. If the results from properly adjusted analyses of nonrandomized cohort data are similar to the results from the analyses of randomized cohort data, then we have some evidence for the generalizability of the randomized cohort results. On the other hand, if important differences do occur, then differences will need to be explained using reasonable arguments not grounded in statistical theory but clinical understanding. Enrollment into the nonrandomized cohort was terminated on February 14, 2013.

Selection of Study Population

Inclusion Criteria:

1. birth weight of $\leq$ 1,000 g,
2. a decision by the attending pediatric surgeon to perform surgery for suspected NEC or IP (The indications for surgery for infants with NEC or IP vary among surgeons and sites),
3. the infant is less than or equal to 8 weeks of age (8 0/7 weeks or less) at the time of eligibility assessment, and
4. patient is at a center able to perform both laparotomy and drainage**.**

Exclusion Criteria:

1. Major anomaly which influences likelihood of developing primary outcome or affects surgical treatment considerations;
2. congenital infection;
3. Prior laparotomy or peritoneal drain placement,

4. Prior NEC or IP,
5. Follow-up unlikely (e.g., mother incarcerated, or currently resides (or plans to move) far from any Network center.), and
6. Infant for whom full support is not provided (including surgical treatment).

Method of Treatment Assignment and Randomization

Infants enrolled into the randomized cohort will be randomized using a variable, permuted, block size scheme. Infants will be randomized by calling RTI International as soon as a decision to perform surgery is made (or via a computer randomization process). Randomized infants will be stratified according to two variables: center and according to the overall risk for death or NDI (higher / lower). The risk stratification formula was primarily developed in the prospective observational model and includes the following variables: birth-weight, gestational age, vasopressor requirement (yes / no), if infant is on high frequency (oscillating or jet) ventilation or not (any modality of conventional ventilation or no mechanical ventilation), pH, FiO2, and preoperative diagnosis (NEC or IP). Treatment beyond the initial surgical management will be unaffected by the trial.

Treatment Masking

As this is a surgical trial, the surgeons will not be masked to randomized treatment. Outcome assessments at 18-22 months will be assessed by evaluators who are masked to the details of the operative intervention(s) that have been performed using standard Network assessments. The statistician is unmasked in this trial.

Sequence of Planned Analyses

Interim Analyses

Annual Data Safety Monitoring Committee (DSMC) meetings will be convened to monitor the progress of the trial and review the accruing safety and efficacy data. Starting one year after at least 11 sites have IRB clearance to start enrolling into the trial, safety data will be monitored. Once follow up data start becoming available (which should occur from the 2nd year onwards) interim efficacy data will also be presented to the DSMC. All figures, summaries, and listings created in the interim analyses and presented to the DSMC will be masked to treatment assignment, but interim analysis results will be unmasked at the DSMC's request.

The treatments in this trial will be conducted during the first few weeks of life for most study subjects, while the primary outcome of death or neurodevelopmental impairment will be assessed at 18-22 months of age. As a consequence, to ensure the safety of participants and trial integrity, interim analysis will be completed using both safety and efficacy data but an emphasis will be placed on adverse events and other safety measures.

The interim analysis plan described below ensures that adverse events are monitored more frequently than measures of efficacy, and the statistical bounds used to detect group differences in adverse event rates are more liberal than those used to determine efficacy. The safety monitoring will consist of summary adverse event tables and a mortality monitoring scheme that uses Pocock type sequential testing boundaries. The scheme for monitoring the primary efficacy variable will use Lan-DeMet's spending functions to implement O'Brien-Fleming type boundaries while accounting for unequally spaced interim analyses. Lan-DeMet's spending functions will be used to estimate the boundaries at each interim efficacy analysis, because we cannot precisely predict beforehand what proportion of trial enrollees will have primary outcome data available before each of the annual DSMC meetings.

Adverse events during the course of the trial treatment period will be prospectively monitored, as will clinical morbidities throughout hospitalization. The enrolled population is extremely high risk, and their hospitalization can produce a great number of expected adverse events and clinical diagnoses. Rates of these events, historically observed among similar extremely low gestation/birth weight infants, will be provided to the DSMC.

The set of clinical outcomes monitored were:

- Death
- Number of surgeries
- Surgical complication
- Parenteral alimentation
- Days on parenteral alimentation
- Days in hospital
- Severe IVH
- PVL
- Seizures
- PDA prior to enrollment
- PDA treated
- EOS
- LOS
- ROP
- BPD

- Death or serious morbidity (Hosp.)
- Death or Severe IVH
- Death or sepsis
- Death or ROP
- Death or BPD
- Causes of Death
    - CNS insult
    - Gastrointestinal
    - Immaturity
    - Infection
    - Malformation
    - Pulmonary
    - Other

For each of the above conditions, summary tables will tabulate counts by treatment group and cohort, but these summaries will not include inferential statistics such as p-values or confidence intervals.

Interim Mortality Assessment

In addition to the annual production of safety summaries suitable for presentation to the DSMC, interim analyses using Pocock type group sequential boundaries will be used to monitor group differences in mortality. For the first 120 randomized enrollees, the DCC would track rates of neonatal mortality after each 30 enrollees reach the Network Status (death, discharge, transfer or 120 days of age). Thereafter, if no trends of differences between groups develop, these comparisons will be done after every 60 enrollees reach Network status. Per the protocol, the computed statistic at each of these safety looks will be compared to Pocock boundaries that are constructed prior to any interim analysis so that an overall alpha level of 5% is maintained. If any trends of differences between groups develop, the DCC would notify the DSMC and present them an unplanned interim analysis report.

The Pocock boundaries for the interim analysis of mortality were calculated by using the gsProbability function in the gsDesign package for R. For boundary calculations, a scenario of six interim looks plus one final analysis was considered. The following code was used:

```
library(gsDesign)
gsProbability(k=7, theta=0, n.I=c(30,60,90,120,180,240,300),
              a=rep(-2.516,7), b=rep(2.516,7))
```

The output from this call to gsProbability was:

```
              Lower bounds   Upper bounds
  Analysis  N    Z   Nominal p  Z   Nominal p
         1  30 -2.52    0.0059 2.52    0.0059
         2  60 -2.52    0.0059 2.52    0.0059
         3  90 -2.52    0.0059 2.52    0.0059
         4 120 -2.52    0.0059 2.52    0.0059
         5 180 -2.52    0.0059 2.52    0.0059
         6 240 -2.52    0.0059 2.52    0.0059
         7 300 -2.52    0.0059 2.52    0.0059

Boundary crossing probabilities and expected sample size assume
any cross stops the trial

Upper boundary
          Analysis
  Theta      1      2      3      4      5      6      7 Total E{N}
      0 0.0059 0.0045 0.0035 0.0029 0.0032 0.0027 0.0023 0.025  291

Lower boundary
          Analysis
  Theta      1      2      3      4      5      6      7 Total
      0 0.0059 0.0045 0.0035 0.0029 0.0032 0.0027 0.0023 0.025
```

The top table of the output just lists the boundary values and the nominal p-values for each boundary value on the scale of a z statistic for sample sizes of 30, 60, 90, 120, 180, 240, and 300 infants. The "Upper boundary" lists the conditional probability of crossing the upper boundary given no previous boundary crossings (upper or lower) under the null hypothesis of no difference. The "Lower boundary" lists the conditional probability of crossing the lower boundary given no previous boundary crossings (upper or lower) under the null hypothesis of no difference. Within each list, the value given in the "Total" column is the overall error attributed to that boundary. Based on these calculations, a boundary set of z statistics equal to {-2.516, 2.516} will produce lower and upper Pocock boundaries for 6 interim analyses and 1 final analysis of mortality.

For this interim assessment of mortality, robust Poisson regression of death at time of Network status will be used as the outcome variable. The robust Poisson regression model will have at least treatment and baseline risk as covariates. In addition, if possible, the robust Poisson regression model will include center as a fixed covariate. This robust Poisson regression model will be called the primary interim mortality model. For the robust Poisson regression, the outcome for the i^{th} infant will be:

$$y_i = \begin{cases} 0 & \text{for survival at Network Status} \\ 1 & \text{for death at Network Status} \end{cases}$$

The covariates will have the form:

$$s_i = \begin{cases} 0 & \text{initial surgery was drain} \\ 1 & \text{initial surgery was laparotomy} \end{cases}$$

$$r_i = \begin{cases} 0 & \text{low baseline risk for death or NDI} \\ 1 & \text{high baseline risk for death or NDI} \end{cases}$$

The model will have the form:

$y_i \mid (s_i, r_i, c_i) \sim Poisson[\lambda(s_i, r_i, c_i)]$, where
$\log[\lambda] = \alpha + \vartheta \cdot s_i + \gamma \cdot r_i + \upsilon_c \cdot c_i$ and c_i represents the site for the ith infant among C total sites.

Once the model is fit, the treatment test-statistic, a z statistic, will be compared to the interval {-2.516, 2.516}. If the test statistic for the treatment effect in the primary interim mortality model is within the {-2.516, 2.516} interval, then per protocol no report will be given to the DSMC. On the other hand, if the test statistic is outside of this interval, then a basic interim mortality analysis report will be produced for the DSMC. The interim mortality analysis report will consist of:

- Treatment effect relative risk ratios from robust Poisson regressions and 95% CIs for all complete infant cohorts (1st 30, 1st 60, 1st 90, etc.).
- Fisher exact test results for comparison of Low/High baseline risk distributions for all complete infant cohorts.
- Frequency tables of mortality by treatment for Low/High baseline risk, with interpretative statements based on the robust Poisson regression model that took into account the treatment by baseline risk interaction.

Any report sent to the DSMC will also contain safety summary tables typically reported in the annual DSMC meetings.

In addition to the above analysis, the interim safety analysis will also include two secondary analyses. First, the observed distributions of High/Low risk within each group will be estimated and compared using Fisher's exact test. Second, a robust Poisson regression of death at time of Network status as a function of treatment, baseline risk, and the interaction between treatment and baseline risk will be completed. This interaction model will have the form:

$y_i \mid (s_i, r_i, c_i) \sim Poisson[\lambda(s_i, r_i, c_i)]$, where
$\log[\lambda] = \alpha + \vartheta \cdot s_i + \gamma \cdot r_i + \gamma_2 \cdot s_i \cdot r_i + \upsilon_c \cdot c_i$.

Interim Efficiency Assessment

To control the Type I error associated with sequential testing of the primary efficacy endpoint, O'Brien-Fleming boundaries will be calculated for three interim and one final analysis. Since we cannot precisely predict beforehand what proportion of trial enrollees will have primary outcome data available before each of these meetings, we will use the Lan DeMets approximation to the O'Brien-Fleming sequential monitoring bounds to account for unequally spaced interim analyses. Original power calculations based on Fisher's exact test indicated that 150 infants per arm would be sufficient for 80% power if 80% and 65% of infants experienced death or NDI in the Drain and Laparotomy arms respectively. Additional simulations were used to estimate power using the following interim analysis plan and a robust Poisson regression model to estimate treatment effect, under the above-mentioned trial assumptions. The estimated power was 82%.

Interim efficacy assessments will be completed at approximately 75, 150, and 225 randomized infants who are at least 22 months corrected age regardless of death. A final analysis will occur at 300 randomized infants and will account for the earlier looks. The following R code was used to predict the O'Brien-Fleming boundaries under these assumptions:

```
library(gsDesign)
OF.design <- gsDesign(k=4,test.type=2,n.I=c(75,150,225,300),
                      sfu="OF")
OF.design$upper$bound
```

In this call to gsDesign(), "k" equals the total number of planned efficacy analyses (interim + final), "test.type" equals 2 so the design is two-sided with symmetrical boundaries, and "sfu" equals OF to indicate O'Brien-Fleming boundaries. Note that only an upper bound spending function was specified since the design is symmetrical. The above call to the gsDesign function resulted in the following boundaries:

Interim Analysis	Lower Boundary	Upper Boundary
1	-4.05	4.05
2	-2.86	2.86
3	-2.34	2.34
4	-2.02	2.02

Since this design is two-sided and not 'closed', it is possible that the study could end without one arm being identified as better than the other. As such, the potential 'beta'

error in this design is the failure to reject a false null hypothesis and not the failure to accept a false null hypothesis.

Because actual enrollment at the time of each interim efficacy analysis will not occur exactly as planned, boundaries will be recalculated using the following R code:

```
OF.update <- gsDesign(k=L, test.type=2, n.I=vec.L, sfu=sfLDOF,
maxn.IPlan=300)
OF.update$upper$bound
```

where *L* equals the total number of completed and still to complete looks such as 4, and *vec.L* contains a modified list of sample sizes at the exact and still to be completed looks such as c(81,150,225,300). Note that the spending function is now "sfLDOF", which specifies that the boundaries will be recalculated given new sample size points based on the Lan-DeMets approximation of O'Brien-Fleming boundaries.

At the time of each interim analysis, the boundaries for the current and future interim analyses will be updated using `gsDesign()` as described above. A robust Poisson regression model will be used to analyze the primary outcome of death or NDI at 18-22 months. For the robust Poisson regression, the outcome for the ith infant will be:

$$y_i = \begin{cases} 0 & \text{for survival without NDI} \\ 1 & \text{for death or survival with NDI} \end{cases}$$

The covariates will have the form:

$$s_i = \begin{cases} 0 & \text{initial surgery was drain} \\ 1 & \text{initial surgery was laparotomy} \end{cases}$$

$$r_i = \begin{cases} 0 & \text{low baseline risk for death or NDI} \\ 1 & \text{high baseline risk for death or NDI} \end{cases}$$

The model will have the form:

$y_i \mid (s_i, r_i, c_i) \sim Poisson[\lambda(s_i, r_i, c_i)]$, where
$\log[\lambda] = \alpha + \vartheta \cdot s_i + \gamma \cdot r_i + \upsilon_c \cdot c_i$ and c_i represents the site for the ith infant among C total sites.

Similar to the safety interim analysis, the test statistic produced for the treatment effect will be a z-statistic. This z-statistic will be compared to the appropriate boundary pair calculated from the updating call to gsDesign(). The actual robust regression modeling at each interim analysis will be completed in SAS, and a figure summarizing the interim

analysis results will be constructed in Excel. Prior interim analysis results will be represented in the figure as well.

Boundary crossings will not require trial stoppage. Any boundary crossing will trigger a review of safety and efficacy data to understand the meaning of the analysis results in light of the uncertainty remaining in the data. The DSMC may choose to continue the trial, regardless of data interpretation. Unless the DSMC requests, the DSMC will remain masked to treatment assignment, but the coordinating center is not masked throughout the trial. Missing data will be identified and summarized; however, no imputation will be performed for the interim analyses.

Interim and Final Analysis Reporting

All final analyses identified in the protocol and in this SAP will be performed only after the last randomized patient has died, withdrawn, or achieved follow-up status. All data checks and cleaning will be completed prior to database lock and completion of the final analyses. In addition, no final analysis will be completed until the NEST database has been locked and this SAP has been approved.

Primary analysis results, baseline measures, and other PI specified collected data within NEST or GDB will be summarized by treatment. These summaries will be made available to the NEST subcommittee following database lock and prior to submission of any initial manuscripts to peer review journals. Any, post-hoc, exploratory analyses completed to support planned study analyses, which were not identified in this SAP, will be documented and reported as instructed by the coordinating center PI. Any results from these unplanned analyses will also be clearly identified in any resulting peer reviewed manuscript as a post-hoc/exploratory analysis.

Final Analyses

Primary Analysis of Randomized Cohort Data

The primary analysis of randomized cohort data will assess the effect of treatment on death or NDI at 18-22 months corrected age, death, and other clinical outcomes as defined in Study Endpoint section above (pages 8-10) and listed on pages 13 and 14. This analysis will be performed using a robust Poisson regression model identical to the model defined on page 17 and 18. If possible center will also be included in the model as a fixed or random effect. The default will be a fixed effect, but a random effect may be used if model convergence issues results from using a fixed effect. The outcome variable will assign one to the event of death or NDI at 18-22 months corrected age and zero otherwise. The treatment variable will assign one to laparotomy and zero to drainage. It follows that relative risks less than one (or less than zero on the log scale) will favor Laparotomy.

Like the efficacy interim analysis, the final test statistic produced for the treatment effect will be a z-statistic. This z-statistic will be compared to the appropriate boundary pair calculated from the updating call to gsDesign() as described in the efficacy interim analysis section. As with the interim efficacy analyses, the actual robust regression modeling at each interim analysis will be completed in SAS, and a figure summarizing the final analysis results will be constructed in Excel. This figure will contain information from all prior interim efficacy analysis results.

In addition to the above formal adjusted analyses, simple unadjusted treatment comparisons of randomized data will be completed, and 95% confidence intervals of differences in proportions will be constructed. The formatting of these tabulations will be determined as the corresponding papers develop.

Secondary Analysis of Primary Outcome Using Randomized Cohort Data

The analyses will include an assessment of whether there is an interaction between treatment (laparotomy or drainage) and disease (NEC or IP). This assessment will use a robust Poisson regression model with will include treatment (laparotomy or drainage), preoperative diagnosis (NEC or IP), and the treatment by preoperative diagnosis interaction. If the Wald test for the interaction term has a p-value less than 0.05, then the conclusion will be the study has produced evidence of an interaction between treatment and pre-operative diagnosis with respect to the primary outcome. If such

evidence is produced, we will produce treatment effect estimates separately for NEC and IP.

The same robust Poisson regression model will be used to compare treatment with respect to death and other secondary outcomes as mentioned on page 20.

Because NDI adjudication is used, analyses of the primary outcome and analysis of any NDI conditional on survival will include sensitivity analyses to determine whether the adjudicated outcomes change any results. Such sensitivity analyses will include removing all babies with adjudicated outcome for the analysis, switching the value of the adjudicated, and if possible, imputing the NDI based on a predictive model of NDI conditional on survival. Since this predictive model will be exploratory, the details of how this predictive model will be build are not defined in this SAP.

Bayesian Analysis of Primary Outcome using Randomized Cohort Data

R and WinBUGS or JAGS software will be used to analyze the final data from a Bayesian perspective. A Bayesian model contains three components: a model or *likelihood* for the data, a *prior* distribution embodying beliefs or historical knowledge regarding the model parameters before new data are observed, and the posterior distribution embodying updated belief probabilities regarding the model parameters and/or functions of the model parameters after the new data are observed.

The outcome from the ith infant used in the model will be:

$$y_i = \begin{cases} 0 & \text{for survival without NDI} \\ 1 & \text{for death or survival with NDI} \end{cases}$$

The covariates will have the form:

$$s_i = \begin{cases} 0 & \text{initial surgery was drain} \\ 1 & \text{initial surgery was laparotomy} \end{cases}$$

$$r_i = \begin{cases} 0 & \text{low baseline risk for death or NDI} \\ 1 & \text{high baseline risk for death or NDI} \end{cases}$$

The model will have the form:

$y_i \mid (s_i, r_i, c_i) \sim Bernoulli[\pi(s_i, r_i, c_i)]$, where
$\log[\pi] = \alpha_c + \vartheta \cdot s_i + \gamma \cdot r_i$ and c represents the site for the ith infant among C total sites.
$c \sim N(0, \sigma_c^2)$

Note that this model differs from the primary analysis model in that it is a log-binomial model and center is included in the model as a random effect.

The Bayesian analysis for this protocol will involve three default priors: "non-informative", skeptical prior, and enthusiastic priors. The "non-informative" prior will represent an individual who would approach the study results with no prior knowledge influencing his/her belief regarding the relative risk of laparotomy vs. drain. On the log relative risk scale, this "non-informative" prior for the treatment effect will have a mean equal to 0 and a variance, σ^2, equal to 10000.

The enthusiastic prior will be based on the effect size used to power the original study. This prior will represent an individual who would approach the study with belief that the true relative risk is 0.8125. This relative risk value is based on the assumption that 80% of infants in the drain group and 65% of infants in the laparotomy group will die or experience NDI by 18-22 months corrected age. Also, the enthusiastic prior will give only a 5% chance that the true relative risk is as high as 1. On the log relative risk scale, this enthusiastic prior will have a mean equal to log(0.8125) and a variance, σ^2 = 0.011223.

The skeptical prior will represent an individual who would approach the study with belief that the true relative risk is one and gives only a 5% chance that the true relative risk is as low as 0.8125 (0.65/0.8 as hypothesized in power calculations). On the log relative risk scale, the skeptical prior will have a mean equal to 0 and a variance, σ^2 = 0.011223.

The final analysis results will include plots of the prior and posterior on the log(RR) and RR scales. Although the most complete inference from each of the above defined prior-to-posterior analyses are the posterior distributions, the posteriors will be summarized by calculating $\Pr(RR \le 1)$ and $\Pr(\log(RR) \le 0)$ using each prior. Also, the posterior distributions for RR will be summarized by calculating equal-tail-area 95% credibility intervals for log(RR) and RR. If $\Pr(RR \le 1) > 0.9$ after using the skeptical prior, then the claim will be that the data contains sufficiently strong evidence against a null hypothesis that it should convince a reasonable skeptic, if the skeptic believes the data to be the output of an unbiased and well conducted study. Similarily, if $\Pr(RR \le 0.8125) < 0.1$ after using the enthusiastic prior, the claim will be that the data contains sufficiently strong evidence against a hypothesis stating the true effect is as large as hypothesized during original sample size calculations.

The model given on the next page will be fit in WinBUGS to estimate the relative risk as well as the uncertainty in the parameter estimates. Note: the model is given in a form that uses non-informative priors, but skeptical and enthusiastic priors will also be applied as defined above. The parameter estimates of this model will be generated by

fitting five MCMC chains. Each chain will involve 10,000 burn-in samples and 200,000 additional samples, which will be thinned by a value of 10. The result will be 100,000 useable samples to estimate the Posterior distributions necessary for making inference about the relative risk.

```
model{
  # subject level Logistic regression likelihood
  for ( i in 1:N){
      y[i]~ dbern( p[i])
      log(p[i]) <- beta1[center[i]] + beta2*trt[i] + beta3*strata[i]
  }
  # center level model (center is a random effect with a normal distribution
  for (j in 1:J){
      beta1[j] ~ dnorm(beta1.hat, tau.beta1)
  }
  beta1.hat ~ dnorm( 0.0, 0.0001) #variance = 1/0.0001 = 10000
  tau.beta1 <- pow(sigma.beta1, -2)
  sigma.beta1 ~ dunif(0, 100)
  # non-informative prior for treatment and baseline risk effects
     beta2~dnorm( 0.0, 0.0001 )  #variance = 1/0.0001 = 10000
     beta3~dnorm( 0.0, 0.0001 )  #variance = 1/0.0001 = 10000

  #Average RR across baseline risk levels.
  logRR <- beta2
  RR <- exp(logRR)
}
```

The above WinBUGS model is a log-binomial regression model. Similar to robust Poisson regression, the log-binomial regression model will correctly estimate the log relative risk for treatment and correctly estimate the standard error for this treatment effect.[5]

Propensity Analysis of Primary Outcome using non-Randomized Cohort Data

The propensity scoring analysis plan for the non-randomized component in NEST will include the following:

1) Prior to analysis, the study statistician, primary clinical investigator, and the coordinating center investigator will select 25 to 50 baseline measures that are at least potentially related to the process of choosing between drain or lap. As such, this selection will not assume the choice depends on any one individual but is the result of many stakeholders participating in the decision.
2) Build a Random Forest (2000 trees) to select a subset of variables that are most highly associated with the choice of lap or drain. Use variable importance measures to determine most important.

[5] Barros AJ, Hirakata VN, Alternative for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio, *BMC Medical Research Methodology*, (2003) Oct 20;3:21. Free text at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC521200/.

3) Using the subset of variables from step 2, build a logistic regression model that predicts initial treatment while accomplishing the following as much as possible:
    1. produces propensity score logit distributions within drain and lap groups that have means close to zero and very similar variances,
    2. with respect to drain and lap groups, produces small propensity score adjusted covariate differences for the overwhelming number of baseline measures identified as relevant (balances covariates among the groups).

The outcome data is not involved in the analysis at this time. Only after a final set of propensity scores are calculated will the outcome data be analyzed while adjusting for propensity scores.

1) Make a yes/no determination if step 3 above gives sufficient evidence that propensity score adjustment can balance relevant baseline covariates among the groups.
2) If a propensity model can be built using covariates that are understood to be related to the treatment selection process and the model produces propensity scores which balance baseline covariate distributions among the groups, then the primary outcome will be analyzed using a robust Poisson regression model that will include propensity score logits as a covariate.
3) If the robust Poisson regression model is fit in step 4, then the estimated risk and propensity score adjusted relative risk will be compared to the risk adjusted relative risk calculated using the randomized subject data.
4) An informal comparison of relative risk point estimates and confidence intervals will be used to assess the similarity of the two relative risk estimates.

Combined Analysis of Randomized and non-Randomized Cohorts

This work is solely exploratory and will only be completed if analysis of non-Randomized Cohort can be considered useful. If the risk adjusted relative risk estimates in the two cohorts are in the same direction, or the propensity adjusted relative risk among the non-randomized cohort is in the same direction as the risk adjusted relative risk estimate in the randomized cohort, then a random-effects meta-analysis will be used to estimate the overall relative risk. The following will be done to complete this meta-analysis:

1) Estimate the overall treatment effect on the log RR scale using a fixed effect model:

$$\log(RR)_F = \frac{\sum w_i[\log(RR)_i]}{\sum w_i},$$

where $w_i = 1/v_i$ and v_i is the variance of the i^{th} estimate of $\log(RR)$.

2) Estimate the Q statistic, which can be used to test for heterogeneity of treatment effect:

$$Q = \sum w_i \left[\log(RR)_i - \log(RR)_F\right]^2.$$

3) Estimate the variance needed for a random effects model:

$$\sigma^2 = \max\left\{0, \frac{Q-(2-1)}{\sum w_i - \left(\sum w_i^2\right)/\sum w_i}\right\}.$$

4) Estimate the random effect weights:

$$\widetilde{w}_i = \frac{1}{\sigma^2 + v_i}.$$

5) Estimate the overall random effect estimate of log(RR):

$$\log(RR)_R = \frac{\sum \widetilde{w}_i \left[\log(RR)_i\right]}{\sum \widetilde{w}_i}.$$

6) Estimate a 95% CI for the random effect estimate of the log(RR):

$$\log(RR)_R \pm \frac{1.96}{\sqrt{\sum \widetilde{w}_i}}.$$

Estimates of baseline risk adjusted relative risk on the log scale and the standard errors will be estimated from the robust Poisson regression models. The variance for the estimated baseline risk adjusted relative risk on the log scale will equal the square of the standard error calculated from the sandwich estimator.

Statistical Appendix A: Exploratory Analysis for Statistician's benefit (will be completed only after all publication required analyses are completed.)

Interestingly, bootstrap analysis of a binary outcome using Poisson regression, when the outcome is a function of treatment, produces an empirical distribution of estimated relative risk on the log scale that is approximated by a normal distribution. Also, an empirical 95% confidence interval for the relative risk on the log scale calculated from these bootstrap samples will approximate a 95% confidence interval for the relative risk on the log scale calculated for a robust Poisson regression model.

For example, using simulated data containing a binary outcome which is a function of treatment and a two level strata variable, the following results occurred:

- The estimated log(RR) using a robust Poisson regression model was -0.209, compared to the true value of -0.208.
- The 95% CI for the treatment effect on the log(RR) scale using robust Poisson regression was -0.347 to -0.071.
- The estimated median bootstrap estimate of log(RR) using Poisson regression was -0.209, based on 10000 bootstrap samples.
- The 95% CI for the treatment effect on the log(RR) scale using the bootstrap estimates was -0.356 to -0.078.

Also, Figure 1 shows that the 10,000 bootstrap estimates of log(RR) from the Poisson regression does approximate a normal distribution.

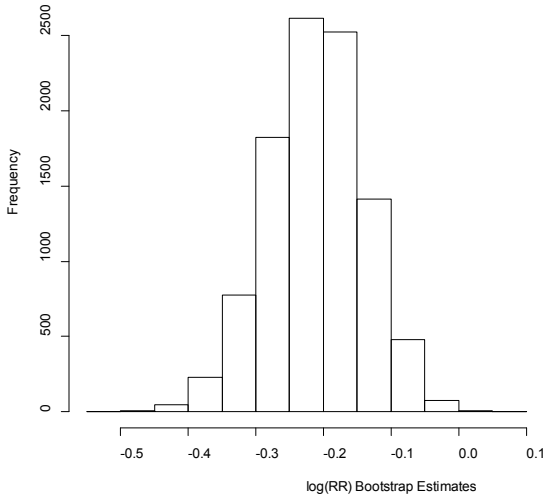

Figure 1

The program used to study the relation between robust Poisson regression and bootstrapped Poisson regression of a binary outcome is included in Statistical Appendix B.

Given the robust Poisson regression estimates of log(RR) are approximately normal, Bayesian analysis of the primary outcome will also use the normal approximation approach of Spiegelhalter, Abrams, and Myles. Namely, the marginal likelihood for the treatment effect will be approximated by a normal distribution with mean equal to the robust Poisson regression estimated log relative risk and standard deviation equal to the sandwich estimate of the log relative risk standard error. The robust Poisson regression model used in this analysis will be the same as the model defined on page 17 and 18. The model will also adjust for center by including center as a fixed effect, if possible.

In addition to assuming the marginal likelihood of the treatment effect is approximated by a normal distribution, normal priors will be used as well. As a consequent, the posterior will be a normal distribution with mean equal to a weighted average of the prior mean and the maximum likelihood estimate of the log relative risk calculated from the data. The variance of the posterior will be the inverse of the sum of the precisions (1/variance).

Prior: $p(\theta) = N(\theta|\mu, \sigma^2)$, where $\theta = \log(\mathrm{RR})$

Likelihood: $p(data|\theta) = N(\hat{\beta}_{trt}, se^2_{sand.})$

Posterior: $$p(\theta|data) = N\left(\frac{\frac{\mu}{\sigma^2}+\frac{\hat{\beta}_{trt}}{se^2_{sand.}}}{\frac{1}{\sigma^2}+\frac{1}{se^2_{sand.}}}, \frac{1}{\frac{1}{\sigma^2}+\frac{1}{se^2_{sand.}}}\right)$$

As with the MCMC analysis using WinBUGS, three default priors will be used, skeptical, "non-informative", and enthusiastic. All three prior-to-posterior analyses will include a plot of the prior, normally approximated likelihood, and posterior. Although the most complete inference from each of the above defined prior-to-posterior analyses are the posterior distributions, the posteriors will be summarized by calculating $\Pr(\theta \le 0)$ for each. Also, the posterior distributions will be summarized by calculating equal-tail-area 95% credibility intervals for θ and for RR=exp(θ). The purpose of analyzing the data using normal approximation methods in addition to the previously defined MCMC methods is to explore the sufficiency of the normal approximation approach for estimating posterior distributions of RR using the robust Poisson regression models. As such, comparisons of the posterior distributions for log(RR) and comparisons of 95% credibility intervals for log(RR) and RR will be made, as well as comparisons of $\Pr(\log(RR) \le 0)$.

Statistical Appendix B: Example of Similarity Between 95% CI of Treatment Effect Using Robust Poisson Regression and Bootstrap Based 95% CI of Treatment Effect Using Poisson Regression.

```
#Next study the distribution of bootstrap estimates of log(RR):
seed.val1 <- 1234
n <- 75
trt.effect <- 0.1875
strat.effect <- 0.06
p.drain <- 0.8
reps2 <- 10000

set.seed(seed.val1)
p.drain1 <- p.drain
p.drain2 <- (1-strat.effect)*p.drain
p.lap1 <- (1-trt.effect)*p.drain1
p.lap2 <- (1-trt.effect)*p.drain2
True.RR <- p.lap1/p.drain1


strat <- c(rep(c(0,1),c(n,n)),rep(c(0,1),c(n,n)))
trt <- rep(c(0,1),c(2*n,2*n))
y <- c(rbinom(n,size=1, p=p.drain1),rbinom(n,size=1,
p=p.drain2),rbinom(n,size=1,p=p.lap1),rbinom(n,size=1,p=p.lap2))
id <- 1:length(y)
sim.dat <- data.frame(id,strat,trt,y)

#Fit the robust Poisson regression model on simulated data.
mod1 <- glm(formula = y ~ trt + strat, data=sim.dat, family=poisson)
mod1.sandwich <- coeftest(mod1, vcov=sandwich)[2,c(1,2)]
mod1.sandwich.CI.lower <- mod1.sandwich[1]-1.96*mod1.sandwich[2]
mod1.sandwich.CI.upper <- mod1.sandwich[1]+1.96*mod1.sandwich[2]
mod1.test.result <- 1*((0 < mod1.sandwich.CI.lower) |
(mod1.sandwich.CI.upper < 0))
CI.diff <- mod1.sandwich.CI.upper - mod1.sandwich.CI.lower
#Construct vector of True RR, Model Estimate, and Sandwich based 95% CI
comparison.vector <- c(log(True.RR), mod1.sandwich[1],mod1.sandwich.CI.lower,
mod1.sandwich.CI.upper, mod1.sandwich[2], CI.diff)
names(comparison.vector) <- c("log(True RR)", "Model Estimate", "95% CI,
Lower", "95% CI, Upper", "SE.trt", "CI.diff")
comparison.vector

# Generate bootstrap estimates of the treatment effect.
# Bootstrap the last simulated dataset, and construct an empirical
distribution of trt effect,
set.seed(seed.val2)
for( i in 1:reps2) {

  ids <- sample(sim.dat[,1], replace=TRUE)
  boot.sim.dat <- sim.dat[ids,]
  boot.sim.dat$id <- rep(1:length(y))
  row.names(boot.sim.dat) <- NULL
  mod.boot <- glm(formula = y ~ trt + strat, data=boot.sim.dat,
family=poisson)
```

```
  mod.boot.coef <- coef(mod.boot)
  if (i == 1) boot.collect <- mod.boot.coef[2]
  if (i > 1) boot.collect <- c(boot.collect,mod.boot.coef[2])

}


print("---------- Look at bootstrap estimates of treatment effect ---------")
print("Estimate of log(RR) from last simulation, based on sandwich var.
est.")
print(comparison.vector)
print("Bootstrap estimate of log(RR)")
print(summary(boot.collect))
print("Bootstrap estimate of 95% CI for log(RR) from last simulation")
print(quantile(boot.collect,c(0.025,0.975)))
windows()
hist(boot.collect, xlab="log(RR) Bootstrap Estimates", main="")
print("----------------------------------------------------------------------")
```