A Randomized Trial of Low-Dose Bevacizumab vs Laser for Type 1 ROP

NCT04634604

Statistical Analysis Plan
October 5, 2021

RETINOPATHY OF PREMATURITY 3

(ROP3)

A RANDOMIZED TRIAL OF LOW-DOSE BEVACIZUMAB VERSUS LASER FOR TYPE 1 RETINOPATHY OF PREMATURITY

Statistical Analysis Plan (version 1.0)

Based on protocol version 1.2

Revision History

VERSION NUMBER		AUTHOR	APPROVER	EFFECTIVE DATE	REVISION DESCRIPTION (INCLUDING SECTIONS REVISED)
SAP	Protocol				
1.0	1.2	R. Henderson	Z. Li	05OCT21	Initial version

Robert Henderson

Digitally signed by Robert Henderson
DN: cn=Robert Henderson, ou=South Wing
Reason:
Location:
Date: 2021-10-05 16:43-04:00

Zhuokai Li
I am digitally signing this document
2021-10-05 17:55-04:00

1.0 Study Overview

Premature infants (birth weight < 1251 grams) with Type 1 ROP in one or both eyes will be randomized 1:1 into the multi-center RCT to determine whether intravitreous bevacizumab (subsequently referred to as BV) injections have a treatment success rate determined at 6-months adjusted age that is non-inferior compared with infants treated with laser photocoagulation (subsequently referred to as LASER). One or two eyes per participant will be eligible for treatment. Randomization will be stratified by zone (I vs II) for the most severe eye and by gestational age (≤ 25 weeks vs > 25 weeks).

Schedule of Study Visits and Procedures

	Enroll	1w*	2w*	4w*	2m*	4m*	6m†	1y†	2y†	3y†	4y†	5y†	6y†
Informed Consent	X												
Medical/Ocular History[a]	X												
Ocular Exam/Classify ROP	X	X	X	X	X	X	X	X	X	X	X	X	X
Retinal Imaging	X[b]												
Cycloplegic Refraction							X	X	X	X	X	X	X
Visual Fix & Follow							X	X					
Visual Acuity									X	X	X	X	X
Visual Fields													X
General Movements Assessment						X[f]							
Bayley-4									X				
IQ test + Neuropsychiatric exam													X
Plasma VEGF	X[c]		X[c]	X[c]		X							
Adverse Events		X[d]	X[d]	X[d]	X[e]	X[e]	X[e]	X[e]	X[e]	X[e]	X[e]	X[e]	X[e]
Systemic Outcomes		X		X			X	X	X	X	X	X	X
Quality of Life								X				X	X

*Exams through 4 months are timed based on initial treatment (and re-treatment when indicated).
†Exams 6 months and beyond are based on adjusted age.
If a vitrectomy or scleral buckle is scheduled prior to the 6-month exam, then the 6-month exam will be completed early and prior to vitrectomy or scleral buckle surgery.

[a] Medical and ocular history to include concomitant medications and pre-existing medical conditions at time of enrollment.
[b] Retinal imaging also done at 40 weeks PMA; and will also be collected when there is treatment failure.
[c] Plasma VEGF testing is optional.
[d] All adverse events recorded at each visit until 4 weeks post-treatment
[e] Only serious adverse events and ocular adverse events recorded beyond 4 weeks post-treatment
[f] BabyMoves General Movements Assessment will be recorded at home by parent(s) at 54 (52-56) weeks PMA.

1.1 Primary outcome Definition – Treatment Success At 6 Months Adjusted Age

The primary objective for the randomized trial is to determine if infants with type 1 ROP treated with intravitreal bevacizumab (subsequently referred to as BV) have a treatment success rate determined at 6 months adjusted age that is non-inferior compared with infants treated with laser photocoagulation (subsequently referred to as LASER).

The primary efficacy outcome will be treatment success at the eye level determined at 6 months adjusted age, where **success is defined as meeting all four of the following criteria** as determined by the investigator:

1. Five to 11 days after treatment (or re-treatment if indicated), there is no increase in the extent or severity of stage 3, or development of new plus disease
2. Twelve to 16 after treatment (or re-treatment if indicated) to 6 months adjusted age, there is no plus disease or severe neovascularization
    - Severe neovascularization is defined as at least one clock hour of extraretinal vascularization that, in the judgment of the investigator, poses a significant risk for retinal detachment and/or dragging.
3. No unfavorable structural outcome at 6 months adjusted age defined as (1) macular ectopia*, (2) a posterior retinal fold involving the macula, (3) a retinal detachment involving the macula, or (4) retrolental tissue or mass obscuring the view of the posterior pole.
    - Macular ectopia must be definite to meet criteria for an unfavorable outcome. Questionable ectopia or vessel straightening alone are not considered unfavorable.
4. No scleral buckle or vitrectomy by 6 months adjusted age.

If any of these criteria are not met, the outcome for that eye is a failure.

With a short time to primary outcome, it is expected that lost to follow-up will be minimal and non-informative. However, there may be a few participants who die or are lost to follow up before the 6-month visit. To use the data that have been collected from these participants before they are lost to follow-up, a time to event analysis will be performed. Censoring will occur at the time of death or lost to follow up before the 6-month outcome. If an eye fails any of the 4 criteria above at one of the visits before the censoring event, that eye will be counted as a failure at that time. The protocol visit date where failure or the censoring event occurs will be used as the timepoint for the analysis. Direct adjustment methods will be used to estimate the success probability to be consistent with the hypothesis below.

1.1.1 Off-protocol treatments and success/Failure

The ROP3 protocol was designed with the input of many clinicians to have protocolized treatments that are the best for patients. However, in individual cases, it is possible that the treating clinician will decide that following the protocol is not the best for that patient.

If off-protocol ROP treatment of the same arm is given before non-response, that eye will be counted as a non-response at that date. For example, if randomized to BV 0.063, if another dose of BV (0.063 or 0.25) is given before non-response criteria is met, that treatment date will be the first non-response date. The reason to consider this a non-response is that any eye treated would probably at least be approaching non-response criteria. A second non-response (as usual) will be considered a failure.

If after a first non-response, an off-protocol ROP treatment is given, the eye will be considered a failure at that time. The reasoning is the same, any eye given a treatment would likely be approaching non-response criteria.
If off-protocol ROP treatment is switching arms (laser to BV, BV to laser), vitrectomy, scleral buckle, or any other off-arm treatment, that eye will be considered a failure at that date. In the protocol, treating with BV after non-response from laser is a failure . Treating the laser arm with BV in the absence of non-response will also be considered a failure. It will be important to correctly identify that the treatment is off protocol. For example, prophylactic laser treatment is a protocol treatment available to the BV arm at 50-65 PMA weeks, or sooner if the infant is being discharged or transferred and waiting to do laser would not be practical. Laser is also allowed as the 2nd treatment if at least 4 weeks have passed, or if there is vitreous organization. In these cases, laser as the 2nd treatment would not be off-protocol, so it would not be considered a failure.

1.2 Primary Statistical Hypotheses

The study is designed as a non-inferiority study to evaluate a 1-sided null hypothesis that the success probability after six months with BV is inferior to LASER by 5% or more in favor of the alternative hypothesis that the success probability with BV is not inferior to LASER (>-5%).

$\mathbf{H_{null}}$ = $BV_{\text{success\% at 6m}}$ minus $\text{LASER}_{\text{success\% at 6m}}$ <= -5% *(BV inferior to LASER)*

$\mathbf{H_{alternative}}$ = $BV_{\text{success\% at 6m}}$ minus $\text{LASER}_{\text{success\% at 6m}}$ > -5% *(BV not inferior to LASER)*

Non-inferiority of BV compared with LASER will be declared if the LOWER limit of the upper 1-sided 97.5% CI for the difference between treatment groups in success proportions (BV minus LASER) is greater than the non-inferiority limit of -5% favoring LASER.

If non-inferiority is declared, a test of no difference (superiority test) for BV compared with laser will be conducted.

1.3 Margin choice

The 5% non-inferiority margin was selected by the planning committee based on clinical judgment. Planning committee members were willing to accept that BV may be slightly inferior to laser (up to 5%) because BV has several other advantages, including ease of administration, less stress to infants during treatment, less myopia, possibly better peripheral vascularization, and possibly better visual fields. Planning committee members were unwilling to accept a margin greater than 5% as "non-inferior," because severe ROP can lead to blindness, and laser is generally successful in preventing blindness.

1.4 Description/ Calculation of Secondary outcomes

1.4.1 Number of re-treatments

The BV group is initially treated with 0.063 mg. If the participant is a non-responder, they are treated with 0.25 mg. Additional treatment after 0.25 mg is at investigator discretion and can potentially be laser treatment.

An eye will be classified as non-response / severe recurrence when any of the following are present:

- Worse ROP 5 to 11 days after treatment, defined as an increase in the extent or severity of stage 3, or development of new plus disease
- No improvement* 12-16 days after treatment (two week exam)
- Beyond the 2 week exam, recurrence of severe neovascularization** or recurrence of plus disease***

* For infants with pre-treatment plus disease, improvement is defined as plus disease no longer being present. For infants with pre-treatment zone I, stage 3, with pre-plus disease, improvement is defined as: (1) pre-plus no longer present (neither plus nor pre-plus disease), or (2) a reduction in severity and/or extent of extraretinal neovascularization. For infants with pre-treatment zone I, stage 3, with neither plus nor pre-plus disease, improvement is defined as a reduction in severity and/or extent of extraretinal neovascularization.

** Severe neovascularization is defined as at least one clock hour of extraretinal vascularization that, in the judgment of the investigator, poses a significant risk for retinal detachment and/or dragging.

*** Recurrent plus disease is defined as *both dilation and tortuosity* meeting or exceeding the degree of abnormality represented by the standard photograph of plus disease. Tortuosity without significant dilation is common after bevacizumab and does not meet criteria for recurrent plus disease.

The laser group can have a second laser retreatment before being released to investigator discretion (which may be BV treatment). The protocol allows retreatments at the initial site or to fill in skip areas. Both of these re-treatments will be counted in the number of re-treatments.

The number of retreatments will be calculated on a per participant and a per eye basis in two different ways

- Until and including the primary outcome.
- Until and include 6 months, and annually thereafter

1.4.2 Refractive error

Refractive error is measured in diopters, to the nearest quarter diopter. The further from zero, the worse the refractive error. Myopia is negative diopters; hyperopia is positive diopters. Normally, babies are mildly hyperopic. Based on previous reports, eyes receiving laser treatment will likely become more myopic than eyes on anti-VEGF treatment. Myopia may develop and progress as much as -9.00 D or greater.

1.4.3 Proportion of study eyes with myopia (spherical equivalent refractive error < -5.0D)

One of the hypothesized benefits of BV over laser is that myopia may be less common. This outcome will be analyzed as a binary outcome in a logistic regression analysis (spherical equivalent refractive error < -5.0D). Separately, myopia values will be tabulated as High Hyperopia (>+5.00D) spherical equivalent. non-myopic (5 - >-0.5 D), some myopia (-0.5 D to =>-5.0D), and severe myopia (< -5.0D).

1.4.4 Bayley Scales of Infant and Toddler Development

The Bayley test has five major component scales that are tested for each child: Cognitive, Language, Motor, Social-Emotional, and Adaptive Behavior. Scores for each scale range from 45 to 155, with a SD of 15. A score of 44 will be assigned to untestable participants[1] (with reason due to developmental delay/cerebral palsy) who do not miss the visit. Other participants who attend the visit but do not complete the questionnaire will be counted as missing.

Bayley composite subscale scores also will be categorized as normal, slightly impaired, or significantly impaired. Participants with scores of 85 or higher are considered normal. Participants with scores of 70 to 84 are considered slightly impaired; and participants with scores <70 are considered significantly impaired. Participants who are untestable are another category below the significantly impaired.

1.4.5 Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R)

The continuous WPPSI-R score has a mean of 100 and a standard deviation of 15. The overall WPPSI-R score has cut points:

- below 70 is Extremely Low
- 70-79 is Borderline
- 80-89 is Low Average
- 90-109 is Average
- 110-119 is High Average
- 120-129 is Superior
- 130+ is Very Superior

1.4.6 Beery–Buktenica Developmental Test of Visual-Motor Integration, 4th edition (VMI-4)

This test assesses the extent to which individuals can integrate their visual and motor abilities. The standardized score has a mean of 100 with a standard deviation of 15.

1.4.7 Visual acuity

Distance visual acuity will be tested in both eyes (when able) beginning at 2 years adjusted age and every year thereafter through age 6 years, beginning with the ATS HOTV testing protocol on the Electronic Visual Acuity Tester (EVA). This procedure provides Snellen equivalent scores that will be converted to the semi-continuous logMAR scale for analyses.

1.4.8 Visual fields

Visual fields will be measured using kinetic sphere perimetry at the 6-year exam. Our measurement is an area measurement. There are 4 areas measured in continuous mean visual field extent (degrees).

Superotemporal (ST)

Superonasal (SN)
Inferonasal (IN)
Inferotemporal (IT)

Blind eyes are scored as 0 in all of the areas. Participants unable to complete the exam will be listed for each treatment group, along with the reason. The score will be calculated in two ways:

1. Complete case
2. Penalizing for non-completion: Participants who cannot complete the exam for developmental reasons will be scored as a zero in all sections. (Participants unable to complete for other reasons will be counted as missing)
   Note: It is possible that BV may cause neurodevelopment problems that may make a participant unable to complete the exam for developmental reasons. If there truly is a developmental effect in the BV group then there would be more participants who would be unable to perform the visual fields test; the second method of scoring will account for this potential difference. If there isn't a developmental effect of BV, then there would be approximately the same number of unable participants in each treatment group. The treatment group comparison would be unbiased but the true estimates within each group would be drawn down artificially compared to the complete case calculation.

1.4.9 Systemic Morbidities

Systemic morbidities will be calculated in three different ways:

- The total number of each type
- The total number of participants (and percentage) having at least one for each type
- A tabulation of the number of events per type per participant

Systemic Outcome Measures and Study Visits at Which They Are Assessed

System	Outcome measure	Study visit				
		Enrollment	1 week after treatment	4 weeks after treatment	6 months adjusted age	Annually 1-6 years adjusted age
General	Surgical wound infection	x	x	x		
General	Delayed healing of surgical wound	x	x	x		
General	Re-hospitalization (defined as inpatient status)				x	x
General	If yes, number of times re-hospitalized				x	x
General	If yes, primary reason(s) for re-hospitalization (from list)				x	x
General	Age at initial hospital discharge (post-birth)			x	x	
General	Weight	x	x	x	x	x
General	Recumbent length / height				x	x
Neurological	Intraventricular hemorrhage (grade)	x				
Neurological	Periventricular leukomalacia	x			x	
Neurological	Hydrocephalus requiring shunt	x			x	
Neurological	Cerebral palsy (mild, mod, severe)				x	x
Neurological	Seizures, excluding febrile	x	x	x	x	x
Neurological	Autism spectrum disorder - physician diagnosed				x	x
Neurological	Cerebrovascular accident / thrombosis	x	x	x		
Neurological	Occipital-frontal circumference (cm)				x	
Pulmonary	Effective inspired oxygen concentration	x	x	x		
Pulmonary	Increased oxygen requirement after treatment		x	x		
Pulmonary	If yes, days until O2 need returns to baseline		x	x		
Pulmonary	Bronchopulmonary dysplasia	x				
Pulmonary	Asthma diagnosed by a physician					x
Pulmonary	Age at which supplemental oxygen discontinued			x	x	x
Pulmonary	Coughing or vomiting blood				x	x
Pulmonary	Need for re-intubation after treatment		x	x		
Renal	Serum creatinine – most recent	x	x			
Renal	Systemic edema	x	x	x		
Renal	Hypertension requiring daily medication	x	x	x	x	x

Hepatic	Scleral icterus	x	x	x		
Hepatic	Serum alanine aminotransferase (ALT) – most recent	x	x			
Hepatic	Serum aspartate aminotransferase (AST) – most recent	x	x			
Hepatic	Serum bilirubin, total and direct – most recent	x	x			
Hepatic	Serum albumin– most recent	x	x			
Gastrointestinal	After tx, time (hours) until full enteral feeds		x	x		
Gastrointestinal	Necrotizing enterocolitis (surgical / nonsurgical)	x	x	x		
Gastrointestinal	Blood in stool, noted grossly	x	x	x	x	x
Cardiovascular	Cardiac arrest		x	x		
Cardiovascular	Vascular thrombosis, not cerebral	x	x	x		
Auditory	Hearing screen at discharge: AABR or OAE; normal / abnormal				x	
Auditory	Hearing aid requirement (right, left, both)				x	x
Auditory	Cochlear implant (right, left, both)				x	x
	Currently using any of the following:					
Pulmonary	Apnea monitor				x	x
Pulmonary	Oxygen				x	x
Pulmonary	Ventilator/CPAP				x	x
Pulmonary	Tracheostomy				x	x
Pulmonary	Pulse oximeter				x	x
Gastrointestinal	Gastrostomy tube and/or tube feeding				x	x

1.4.10 Extent of retinal vascularization

The extent of retinal vascularization will be classified at 1 year as:

- Posterior Zone I
- Anterior Zone I
- Posterior Zone II
- Mid Zone II
- Anterior Zone II
- Zone III
- Full Vascularization

1.4.11 PedEyeQ

Rasch scores for each questionnaire item will be obtained based on response options from published look-up tables available at www.pedig.net, and used to calculate a score for each participant and a treatment group mean for each of the following domains of the Child, Proxy, and Parent PedEyeQ at randomization and at each visit. Domain scores will also be converted to a 0-100 scale to aid in interpretation.

- Child PedEyeQ
  - Functional Vision
  - Bothered by Eyes and Vision
  - Social
  - Frustration / Worry
- Proxy PedEyeQ
  - Functional Vision
  - Bothered by Eyes and Vision
  - Social
  - Frustration / Worry
  - Eyecare
- Parent PedEyeQ
  - Impact on Parent and Family
  - Worry about Child's Eye Condition
  - Worry about Self-perception and Interactions
  - Worry about Functional Vision

1.4.12 BabyMoves

Videos will be uploaded to a REDCap server at Nationwide Children's Hospital (Columbus, OH), and a certified GMA assessment expert will review the video and grade the infant's movements according to Prechtl's GMA as:

- Normal: fidgety general movements are intermittently or continuously present
- Absent: fidgety general movements are not observed or are sporadically present
- Abnormal: fidgety general movements are exaggerated in speed and amplitude

1.4.13 Fix and Follow

Before visual acuity can be assessed with HOTV, we will use "Fix and Follow":

- Assessment of Vision on an eye and participant level:
  - The % of eyes able to fix and follow during fixation preference testing
    - Categories: None, Central, Steady, Maintained
- Assessment of Amblyopia at participant level:
  - This will be based on fixation preference testing results; the distribution of responses will be tabulated as:
    - Present OD, Present OS, Not Present

We will also do a version of these analyses tabulating by if the eye was treated.

1. 5 Sample Size

Based on previous data from the ROP1 dose de-escalation study analyzed using a Bayesian statistical model, the treatment success rate was estimated to be 95% for eyes initially treated with BV 0.063 mg (and retreated with BV 0.25 mg if necessary), and 85% for eyes[2] initially treated with LASER (and retreated with LASER if necessary).

It is expected that there will be a high correlation with respect to success between eyes of the same participant in those with bilateral treatment. Therefore, sample size calculations have been calculated based upon the number of participants to be conservative. The primary analysis will be a time to event analysis. To be conservative again, the power calculations were based on logistic regression.

Assuming that the true difference in success rates is 10% favoring BV, a total sample size of 180 participants (90 per group) will have 90% power to reject a 1-sided null hypothesis that the success rate after six months with BV is inferior to LASER by 5% or more in favor of the alternative hypothesis that the success rate with BV is not inferior to LASER (> -5%), with a type 1 error rate of 2.5%.

The percentage of participants lost to follow-up or withdrawn prior to six months adjusted age is expected to be no more than 15% and equal between the two treatment groups. After adjusting for 15% lost to follow-up, the total sample size is 212 participants (~106 per group).

1.6 Classifying Visits

Data will be collected on the day of treatment with BV or laser. Post-treatment study exams will occur at 1, 2, and 4 weeks, and at 2 and 4 months after initial treatment (and re-treatment if indicated).

Additional study exams will occur at adjusted ages 6 months, 1, 2, 3, 4, 5 and 6 years (calculated as chronological age minus number of weeks born prematurely). It is anticipated that non-study visits will occur more frequently, and the timing of these will be at investigator discretion.

Protocol-specified study visits (and target, allowable windows) are as follows:

Protocol Visit Schedule

Visit	Target Day	Target Window	Allowable Window

		(Around Target Day)	(Around Target Day)
ROP diagnosis Enrollment / Randomization	N/A	N/A	N/A
Randomized Treatment (with Injection or Laser)	Day of Randomization	+ 1 day	+ 2 days
1 week post-treatment*	Treatment Date + 1 week	5 to 9 days	5 to 11 days
2 weeks post-treatment*	Treatment Date + 2 weeks	12 to 16 days	12 to 16 days
4 weeks post-treatment*	Treatment Date + 4 weeks	25 to 31 days	17 to 38 days
2 months post-treatment*	Treatment Date + 2 months	54 to 68 days	39 to 83 days
4 months post-treatment*	Treatment Date + 4 months	115 to 129 days	84 to 168 days
6 months corrected age	EDC + 6 months	169 to 197 days	124 to 242 days
1 year corrected age	EDC + 1 year	335 to 395 days	243 to 547 days
2 years corrected age	EDC + 2 years	700 to 760 days	548 to 912 days
3 years corrected age	EDC + 3 years	1065 to 1125 days	913 to 1277 days
4 years corrected age	EDC + 4 years	1431 to 1491 days	1278 to 1643 days
5 years corrected age	EDC + 5 years	1796 to 1856 days	1644 to 2008 days
6 years corrected age	EDC + 6 years	2162 to 2222 days	2009 to 2375 days

* Post-initial treatment. If re-treatment when indicated, exams will be repeated after re-treatment.

Due to retreatments, it is possible that retreatment windows will overlap with the corrected age visits. In these cases, the corrected age visits will take priority.

The goal is for all participants to complete all scheduled visits. However, participants who (because of unforeseen circumstances) are unable or unwilling to return for all follow-up visits will be permitted to return for as many study visits as possible as an alternative to withdrawal from the study. When a participant is placed into this status, missed visits will not be recorded as protocol deviations (since they would not be recorded as protocol deviations if the participant was dropped from the study).

If any study-mandated examination is deferred because of an infant's unstable medical status, then that examination will be done as soon as possible within the allowable window.

1.7 Planned Interim Analyses

There will be no formal guidelines for stopping for futility or efficacy, or re-estimation of sample size. The neurodevelopment outcomes in the later years of the study are an essential outcome.

2.0 Description of Statistical Methods

2.1 General Approach

Analysis methods include survival analyses, generalized linear models, tabulations, and plots. Assumptions will be checked.

The primary analysis will follow an intent-to-treat (ITT) principle.

There are sensitivity analyses for the primary analysis.

There are many outcomes and timepoints; we have a multiple comparison plan.

2.2 Analysis datasets

2.2.1 Primary analysis (Eye Level)

The primary efficacy analysis will follow an intent-to-treat (ITT) principle. For the primary analysis, there will be no imputation of data for participants who die or are lost to follow-up prior to the six-month adjusted-age outcome visit. However, participants' eyes can reach failure before the 6-month visit, and they will be included as failures.

If a fellow eye becomes eligible after the first eye, the protocol allows this eye to be treated. However, these eyes will not be included in the analysis dataset to avoid biasing the treatment comparison. Imagine the scenario where the first eye is treated with BV, and then weeks later the second eye becomes eligible. This second eye may already be resistant to BV's benefits. There is not the same sort of effect to worry about in the laser group.

2.2.2 Complete Case Datasets

Complete case datasets on both a participant level and eye level basis will be used for secondary outcomes and safety analyses.

There is a contingency plan for the primary analysis if Cox model assumptions fail (section 2.4.4) that would use a binary success/failure outcome for each participant. If a participant has one study eye, that participant is counted as a failure if the study eye fails. If a participant has two study eyes, a failure in either will count the participant as a failure.

2.2.3 Multiple Imputation for secondary outcomes

Multiple imputation datasets will be created for continuous FDR-controlled secondary outcomes on a yearly basis (see section 2.5). There will be participant and eye level versions depending on the outcome. The imputation models will use the dependent variable, zone (I vs II) for the most severe eye at randomization, continuous gestational age at randomization, and treatment group.

There are no categorical secondary outcomes that will use imputation. As the continuous outcomes with potential missingness are assumed to be normal, Monte Carlo Markov Chain (MCMC) full-data imputation will be used. This method allows for non-monotone data imputation.

If the outcome is measured at multiple timepoints, the values of the outcome at previous timepoints will be included in the imputation model.

We will use 200 burn in iterations and have 100 imputations. A seed will be used for reproducibility. If a participant has died before the analysis timepoint, they will be removed from the dataset.

> Note: There are some outcomes that are measured at multiple timepoints (Refractive error, Visual Acuity, PedEyeQ), so likelihood-based methods were considered to handle missingness. However, these outcomes are not measured at baseline, so it would be possible that missing participants after baseline would not be able to contribute information.

2.2.4 Multiple Imputation for primary analysis contingency plan dataset

There is a contingency plan for the primary analysis if Cox model assumptions fail (section 2.4.4) that involves the multiple imputation of a binary success/failure outcome for each participant. If a participant has one study eye, that participant is counted as a failure if the study eye fails. If a participant has two study eyes, a failure in either will count the participant as a failure.

A binary variable cannot meet the multivariate normality assumption for MCMC imputation, so the fully conditional specification method will be used. Multiple imputation will be used for binary success/failure on participants lost to

follow up for reasons other than death. The imputation models will use zone (I vs II) for the most severe eye, continuous gestational age at randomization, treatment group, and the dependent variable (in that order).

We will use 200 burn in iterations and have 100 imputations. A seed will be used for reproducibility. If a participant has died before the analysis timepoint, they will be removed from the dataset.

2.3 Multiplicity

Type I error for the primary outcome analysis will be 5%, as only one analysis is pre-specified as primary. All other outcomes and their analyses will be pre-specified as either secondary or exploratory. Multiplicity in pre-specified secondary outcome analyses will be controlled using the False Discovery Rate (FDR) method. We plan to report results for secondary outcomes annually; the FDR will be set to 5% for each of the annual reports and will apply to all pre-specified secondary analyses.

For FDR-controlled outcomes, an estimate and FDR-adjusted confidence interval for the difference between treatments will be provided. Estimates and 95% confidence intervals within each treatment also will be provided for all secondary outcomes.

For some outcomes, the continuous analysis is controlled with FDR and the categorical analysis is exploratory. Continuous outcomes have more power, and the categorical outcomes will be used to help with interpretation.

2.4 Primary Outcome

2.4.1 Primary analysis

The proportion of study eyes that achieve success will be tabulated by treatment group. The number of events, censorings, and 6-month visits will also be tabulated by treatment group. The denominator will be the number of eyes that have been randomized in the study. A tabulation of events with the denominator as the number of eyes that have been randomized in the study without deaths and lost to follow-up will also be displayed. Summary statistics for survival time accounting for censoring will be reported.

The hazard ratio of failure for BV: Laser and a one-sided upper tailed confidence (CI) will be calculated with a Cox model with robust variance estimation to control for correlation arising from participants contributing two study eyes to the analysis. The model will include treatment group as the independent variable of interest and adjust for the stratification factors of zone (I vs II) for the most severe eye and continuous gestational age at randomization. The probability of treatment success at 6 months adjusted age will be estimated for each treatment group using the direct adjustment method. The difference in the success probability at 6 months with the upper 1-sided 97.5% CI will be calculated based on a normal distribution.

<u>Assumptions</u>
Functional form of continuous gestational age at randomization will be assessed by plotting the observed distribution of cumulative sums of martingale residuals (Y axis) vs the values of gestational age (X axis). Twenty simulated sets of residuals with correctly specified models will be plotted over our residuals; a seed will be used for reproducibility. If our residuals are very different from the simulated model, this suggests misspecification of the functional form. A Kolmogorov-Type supremum test will be conducted comparing our residuals to 1000 random simulations. A significant result on this test suggests misspecification. Methods that may be used for solving misspecification include transforming or categorizing gestational age at randomization.

Proportional hazards will be checked separately for continuous gestational age at randomization, zone (I vs II) at randomization, and treatment group. Plots will be created with the observed distribution of cumulative sums of martingale residuals (Y axis) vs the length of follow-up (X axis). Twenty simulated sets of residuals with truly proportional hazards will be plotted over our residuals; a seed will be used for reproducibility. If our residuals are very different from the simulated model, this suggests non-proportional hazards. A Kolmogorov-Type supremum test will be conducted comparing our residuals to 1000 random simulations. A significant result on this test suggests non-proportional hazards. A Kaplan-Meier plot will be generated and examined for proportional hazards violations.

If the proportional hazards assumption fails, a model with an interaction between treatment group and a function of time may be an option.

Intermittent missingness is expected to be rare for the primary outcome. These participants are followed very closely as there is risk for blindness.

2.4.2 Tipping Point Analysis for Imaging

To mitigate unmasked assessment bias, wide-angle retinal images will be collected when treatment failure is declared. Images will later be reviewed by masked expert readers to define success or failure. If the site does not have the hardware for wide-angle retinal imaging, a second ocular exam (by a different examiner) will be used in place of the masked image analysis. The number of discrepancies between the site diagnosis and the masked reviewer diagnosis will be calculated. If there are cases where the investigator declared failure and masked review of the images does not confirm failure, then a sensitivity tipping point analysis will be done to evaluate the number of discrepancies required to tip results from significant to not or vice versa. Depending on the numbers of disagreements, we will tabulate either all tipping points or summarize by every few cases.

For each treatment group we will create a table like the one below:

Failures tipped to successes	New success probability difference	New Lower CI Probability Difference (-5% is the NI limit)

2.4.3 Deaths as failures/ Potential competing risks

To check for informative censoring, a sensitivity analysis will be performed where deaths before failure are counted as failures instead of censoring. If deaths are believed to be independent of treatment, this will overestimate the failure rate in each group.

However, if one group's failure rate meaningfully increases compared to the other, this may indicate informative censoring. In this case, the primary analysis may be adjusted to use Fine and Gray's proportional sub-distribution hazards model to fit deaths as a competing risk to protocol treatment failure.

2.4.4 Primary outcome contingency plan

If Cox model assumptions fail and cannot be remedied, Barnard's exact test will be performed on the complete case participant level success/failure outcome described in section 2.2.2. The exact upper 1-sided 97.5% CI for the risk difference will be obtained.

The contingency model will also be put through sensitivity analyses described in 2.4.2 and 2.4.3. As a further sensitivity analysis, a dataset with the binary failure imputed will also be used. Details for imputation are in section 2.2.4.

2.5. Secondary Outcomes

<u>General</u>

We will tabulate or provide summary statistics on all the secondary outcomes overall and by treatment group at each timepoint they are analyzed.

Many outcomes will be analyzed using some type of linear model (see the table and subsection below). For each of these outcomes, the methods, assumptions and diagnostics process are the same (below).

The hypotheses for these outcomes are

H_0: treatment group difference = 0

H_A: treatment group difference $\neq 0$

Exceptions to these rules will be noted in subsections.

Regression Methods

Binary outcomes will be analyzed using Poisson regression (log link) with generalized estimation equations (exchangeable structure) to control for correlation arising from participants contributing two study eyes to the analysis (if applicable to the outcome). The coefficients, when exponentiated, will estimate relative risk. If the deviance/DF is <0.8, the model is underdispersed and logistic regression will be used to estimate odds ratios.

Continuous outcomes will be analyzed using linear regression with generalized estimating equations (exchangeable structure) to control for correlation arising from participants contributing two study eyes to the analysis (if applicable to the outcome). Depending on the empirical distribution, a normal, Poisson, or negative binomial distribution for the model may be used.

Independent covariate adjustments are treatment group, zone (I vs II) for the most severe eye, and continuous gestational age.

For analyses on an eye level, an exchangeable correlation matrix will be used to account for correlation between eyes for participants who contribute two eyes to the study.

> Note: Longitudinal/ Multi-Level Models were considered, but it is anticipated that investigators will want to publish on yearly endpoints rather than longitudinal models.

Checking Assumptions

Standard residual diagnostics will be performed for all analyses. The linearity (or log linearity) assumption of covariates will be evaluated using descriptive scatterplots and by categorizing each of the baseline factors in the model to check for approximate linearity (or log linearity) of the coefficients across ordered categories.

For continuous outcomes, potential outliers will be identified, and a sensitivity analysis will be performed to evaluate the effect of these outliers on the primary outcome results If values are highly skewed, then we will consider categorization, transformations, worst-rank regression, MM estimation methods or non-parametric randomization-based method of Koch et al.

For binary outcomes, if the Poisson regression model fails to converge, we reserve the option to use exact logistic regression or Barnard's test to compare treatment groups. We may also choose to use Barnard's test if there are less than 10 outcome events per covariate, which is considered too small for reliable estimation. Barnard's test does not allow for covariate adjustment.

If results differ between models with violated assumptions and models with the changes for robustness, the robust method will be used for trial conclusions and formulation of resulting recommendations.

Patterns of missingness will be analyzed. Suspected informative missingness will be noted and results will be interpreted cautiously.

Sensitivity analysis for the FDR controlled analysis – Multiple imputation

Multiple imputation for missing data (other than death) will be performed as a secondary approach for all FDR controlled outcomes (section 2.2.3). Results of the analysis with imputation of missing data (other than death) will be assessed for consistency with the primary analysis. If the ITT analyses with no imputation and analyses with imputation give inconsistent results, exploratory analyses will be performed to evaluate possible factors contributing to the difference.

517

	6-Month	1 year	2 year	3 year	4 year	5 year	6 year	FDR or Exploratory	Outcome type	Additional Analyses
Number of re-treatments	X							Exploratory	Count	
Refractive error	X	X	X	X	X	X	X	FDR	Continuous	Regression
Proportion of study eyes with high myopia (spherical equivalent refractive error <= -5.0D)	X	X	X	X	X	X	X	FDR	Binary	Regression
Bayley Neuro-development continuous[1]			X					FDR	Ordinal analyzed as continuous	Regression
Bayley Neuro-development categories[1]			X					Exploratory	Categorical	See section 2.5.1
Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R), continuous							X	FDR	Ordinal analyzed as continuous	Regression
Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R), categorical							X	Exploratory	Categorical	
Beery–Buktenic Developmental Test of Visual-Motor Integration, 4th edition (VMI-4)							X	FDR	Ordinal analyzed as continuous	Regression
Fix & Follow	X	X	X	X	X	X	X	Exploratory	Categorical	
Visual acuity			X	X	X	X	X	FDR	Continuous	Regression
Visual fields							X	FDR	Continuous	Regression
Systemic morbidities	X	X	X	X	X	X	X	Exploratory	Counts	
Extent of retinal vascularization		X						Exploratory	Categorical	
Proxy PedEyeQ[2]		X				X	X	FDR	Ordinal analyzed as continuous	Regression
Parent PedEyeQ[3]		X				X	X	FDR	Ordinal analyzed as continuous	Regression
Child PedEyeQ[4]						X	X	FDR	Ordinal analyzed as continuous	Regression
BabyMoves	X[5]							Exploratory	Categorical	See section 2.5.2

518

1. Contains
    a. Bayley Cognitive subscale
    b. Bayley Language subscale
    c. Bayley Motor subscale
    d. Bayley Social-Emotional subscale
    e. Bayley Adaptive Behavior subscale
2. Contains separate domains:
    a. Functional Vision
    b. Bothered by Eyes and Vision
    c. Social
    d. Frustration / Worry
    e. Eyecare
3. Contains separate domains
    a. Impact on Parent and Family
    b. Worry about Child's Eye Condition
    c. Worry about Self-perception and Interactions
    d. Worry about Functional Vision
4. Contains separate domains:
    a. Functional Vision
    b. Bothered by Eyes and Vision
    c. Social
    d. Frustration / Worry
5. BabyMoves is done at 4 months

2.5.1 Bayley Neurodevelopmental Outcomes

Continuous Analysis

The neurodevelopment analysis will follow a modified intent-to-treat (ITT) principle. Data will be included only from participants who complete the outcome at the time point in question. There will be no imputation of data for participants who are lost to follow-up or who die prior to the Bayley testing. Assumptions will be checked with methods described in section 2.5.

Each of the five domains will be analyzed as a dependent variable using linear regression with generalized estimation equations (exchangeable structure) to control for correlation arising from participants contributing two study eyes to the analysis (if applicable to the outcome).

If distributional assumptions cannot be met with a transformation, Worst-rank regression will be used. Ranks can be calculated from the assigned values on the outcomes as informative missingness from developmental problems are going to be assigned a score of one worse than the lowest possible score on the test as described in section 1.4.

As with other FDR controlled outcomes, a version of the analyses using multiple imputation will be completed as a sensitivity analysis. If a non-parametric method is used, multiple imputation may be inappropriate due to violation of distributional assumptions. In this case, we will use inverse probability weighting to deal with missing data as a sensitivity analysis. To predict missingness, we will use the zone (I vs II) for the most severe eye at randomization, continuous gestational age at randomization, birth weight in grams, and treatment group.

Exploratory Categorical Analysis

For the categorical analysis, the proportion of participants within each score category at 2 years adjusted-age visits will be compared to evaluate whether there is a difference in this outcome between randomized treatment groups. Depending on cell counts in each treatment group and the Bayley category (normal, slightly impaired, severely impaired), an ordinal logistic regression model will be considered. If counts are too small or the proportional odds assumption fails, the outcome will be dichotomized at <70 vs ≥ 70 for significantly impaired vs not. Proportional odds are a strong assumption, so this fallback plan is likely. There may also be an additional model with a cut point at <85 vs ≥ 85 for non-normal vs normal. There will not be a sensitivity analysis for the exploratory categorical analysis as its purpose is to aid in interpreting the continuous analysis.

2.5.2 Exploratory BabyMoves Analysis

4 months

In addition to tabulations, Fisher's exact test will be used to generate a p-value comparing the proportion in each category by treatment group.

Relationship with later neurological assessments

Tabulations of 4-month BabyMoves versus the Bayley and Wechsler categories at each timepoint will be displayed overall and by treatment group.

For each timepoint of continuous Bayley scores, WPPSI-R, and VMI-4, raw correlations with BabyMoves category from the 4-month assessment will be calculated.

There will also be regression models fit for Bayley, WPPSI-R, and VMI-4 scores.

- Univariable models with independent variables of treatment group, zone (I vs II) for the most severe eye at randomization, continuous gestational age, and BabyMoves (GMA) category (normal, absent, abnormal)
    - Dependent variables will be Bayley, WPPSI-R, and VMI-4 scores separately
- Three separate multivariable regression model with dependent variables of Bayley, WPPSI-R, and VMI-4.
    - BabyMoves (GMA) category (normal, absent, abnormal), will be the independent variable of interest. Adjustment covariates of treatment group, zone (I vs II) for the most severe eye at randomization, and continuous gestational age will be considered.
    - Multicollinearity will be assessed. It is possible that treatment and other baseline covariates are associated with the BabyMoves scores so covariate adjustments may "adjust away" any relationship between the GMA and the other neurological assessment. If this is the case, the variable that causes the relationship to "adjust away" will be discussed.
    - If a potential treatment interaction is indicated, we will also display models for each treatment group separately.

2.6 Baseline Descriptive Statistics

For all randomized participants, baseline characteristics will be tabulated. The baseline characteristics will be tabulated according to randomized treatment group, overall, and by completion (vs non-completion) of each follow-up visit. Baseline characteristics will also be tabulated according to mortality.

- Subject level characteristics:
    - Distribution of race and gender
    - The median, quartiles, and range will be calculated for the following data which are not expected to be normally distributed:
        - Birthweight in grams
        - Gestational age in weeks
        - Birth age at time of injection in weeks and days

- Eye level characteristics:
    - Distribution of examination findings in the study eye will be categorized as:
        - Zone 1 ROP with plus disease
        - Zone 1, Stage 3, without plus disease
        - Zone II, Stage 2, with plus disease
        - Zone II, Stage 3, with plus disease

2.7 Subgroup Analyses

The following subgroup analyses will be considered:

- Sex (male vs female)
- Race/ethnicity (white non-Hispanic vs non-white or Hispanic)
- Zone I vs Zone II ROP
- ROP subtype (Zone II Stage 2 with plus, Zone II, Stage 3 with plus, Zone I any ROP with plus, Zone I stage 3 without plus)
- Gestational age (≤25 weeks vs. >25 weeks)

The approach for these analyses will be to estimate the treatment effect and 95% confidence interval within each subgroup using the same analytic methods to the primary analysis. That is, the subgroup analysis will be performed by including the main effect of the subgroup factor and its interaction effect with the treatment group in the model for the primary analysis. If a potential treatment interaction is indicated, we will also display models for each treatment group separately.

Interpretation of the results will depend on whether the overall analysis demonstrates a significant treatment group difference; in the absence of an overall difference, subgroup analysis results will be interpreted with caution. We will consider the clinical significance of this interaction as well as the effect on the treatment coefficient and AIC. For example, if the interaction effect size is very small (clinically insignificant), but precise (statistically significant), it may be an artifact of modeling/overfitting and clinically misleading to describe. Sometimes this happens when there is some sort of relationship between an unspecified effect and predictors.

2.8.1 Safety Analyses

A tabulation of all study eye ocular, non-study eye ocular, and systemic adverse events will be tabulated according to treatment groups.

Ocular (study eye-level)

Systemic adverse events will be reported with treatment grouped two ways:

The first grouping will be according to randomized treatment assignment (ITT groups). The second grouping will be based upon how the participants were actually treated (or retreated), with eyes that ever received bevacizumab included in one group and eyes that never received bevacizumab included in the other, regardless of randomized treatment assignment.

Within the two groups above (i.e., ITT and treatment received), ocular adverse events will be tabulated separately for the two treatment groups and separately for non-study eyes. The total number of adverse ocular events occurring in each group will be counted. For outcomes that are possible in each treatment group, the percentage of eyes experiencing each outcome at least once will be compared between treatment groups, separately for each outcome.

If there are more than 12 events, a logistic regression model will be used to generate P-values to compare events between treatment groups. An exchangeable correlation matrix will be used to account for correlation between eyes for participants who contribute two eyes to the study. For outcomes with no more than 12 events, Barnard's exact test will be used to compare the outcomes between the treatment groups without adjustment for correlation between eyes.

The following ocular adverse events are of primary interest:

- Endophthalmitis
- Retinal detachment
- Cataract
- Vitreous hemorrhage
- Neovascular glaucoma

- Iris neovascularization

Systemic (participant level)

Systemic adverse events will be reported with treatment grouped two ways:

The first grouping will be according to laterality and randomized treatment assignment (ITT groups):

1. Unilateral participants randomized to laser
2. Bilateral participants randomized to laser
3. Unilateral participants randomized to bevacizumab
4. Bilateral participants randomized to bevacizumab

The second grouping will be classified based upon how they were actually treated (or retreated), with infants who ever received bevacizumab included in one group and infants who never received bevacizumab included in the other, regardless of randomized treatment assignment.

The following systemic adverse events are of primary interest:

- Death
- Serious adverse event (at least one)
- Hospitalization (at least one)
- Secondary systemic adverse events of interest:
  - Delayed healing of surgical wound
  - Cerebrovascular accident / thrombosis
  - Necrotizing enterocolitis (surgical / nonsurgical)

Barnard's exact test will be used to generate P-values to compare events between treatment groups.

2.9 Protocol Adherence and Retention

Protocol deviations will be tabulated by randomized treatment group.

Visit and call completion will be summarized by randomized treatment group. A flow chart will account for all participant visits and phone calls.

References

1. Adams-Chapman I, Heyne RJ, DeMauro SB, et al. Neurodevelopmental impairment among extremely preterm infants in the neonatal research network. *Pediatrics.* 2018;141(5).
2. Stahl A, Lepore D, Fielder A, et al. Ranibizumab versus laser therapy for the treatment of very low birthweight infants with retinopathy of prematurity (RAINBOW): an open-label randomised controlled trial. *Lancet.* 2019;394(10208):1551-1559.