INTERMITTENT EXOTROPIA STUDY 5 (IXT5)

A Randomized Clinical Trial of Overminus Spectacle Therapy for Intermittent Exotropia

STATISTICAL ANALYSIS PLAN

Version 2.2
February 25, 2021

Based on Protocol Version 3.0 (November 22, 2019)

21 Revision History

VERSION NUMBER		AUTHOR	APPROVER	EFFECTIVE DATE	REVISION DESCRIPTION (INCLUDING SECTIONS REVISED)
SAP	Protocol				
1.0	V2.0 (Nov 16, 2016)	Rui Wu	Michele Melia	April 2, 2018	Initial version (Completed prior to first DSMC review of study data)
1.1	V2.0 (Nov 16, 2016)	Danielle Chandler			**Section 4:** To reflect the stated time-to-event analysis, changed name of deterioration outcomes to be "deterioration by" the 12-month and 18-month timepoints instead of "deterioration at" these timepoints, **Section 5.1**: On outcome evaluating "no spontaneous tropia," added acknowledgement that some participants may have met this criteria at baseline if they had a baseline control score or 2 or better (≤2) and did not have a spontaneous tropia at any time during the baseline visit at distance or at near. **Section 7.4:** The comparison of refractive error between treatment groups will be performed at the 12-month visit only, so removed erroneous reference to an 18-month comparison. **Section 7.6:** Added this new section describing tabulations on the IXT Symptom Survey. **Section 8.1:** Added accommodative convergence/accommodation (AC/A) ratio as a baseline subgroup factor of interest. Also added a paragraph on the rationale for each baseline subgroup factor, and expected direction of effect for some of the factors. **Section 9.2:** On the assessment of esodeviation, changed to assess "esodeviation by PACT" instead of "tropia by SPCT."
2.0	V3.0 (Nov 22, 2019)	Amra Hercinovic	Michele Melia	March 16, 2020	**Re-ordered analyses and created new headings throughout SAP.** **Section 1.0:** Clarified that all references to refractive error refer to refractive error in the more myopic eye. Section 2.1: Edited to explain why ADHD will not be included in primary analysis.

					Section 2.1.7: Added sensitivity analysis #5 that will include additional adjustment factors. Section 2.1.9: Addressed how imputation will be handled under Protocol Amendment II for participants who missed certain visits. Section 2.2: Explained that analysis will be repeated in a cohort limited to those who were prescribed weaning before Protocol Amendment II. Section 2.3: Revised to say that effect of weaning will only be calculated in the overminus group and will not use ANCOVA but rather a t-test or Wilcoxon depending on normality of distribution. Also edited to say that the analysis will be limited to those who were prescribed weaning, and deleted part about Bonferroni correction as it will not be done here. Also added paragraph on calculation of the retention of 12-month treatment effect at 18 months. Section 3.1: Edited to say that treatment groups will be compared using a Z-test as long as one of the proportions is between 0.05 and 0.95. Edited adjustment factors, deleting distance PACT and leaving only baseline factors of the deterioration clinical outcomes (stereo and control). Added paragraph on how proportional hazards assumption will be checked. Section 3.2: Edited to say that cause-specific deterioration will be calculated only if overall deterioration is significant. Revised adjustment factors as in Section 3.1. Explained that alpha will be split between time points and gatekeeping strategy will be used. Stated that participants with nil stereo at baseline will be excluded. Deleted sensitivity analysis on deterioration outcome (originally Section 4.3). Section 4: Stated what statistics will be tabulated/calculated for all outcomes.

					Explained that FDR multiplicity adjustment will be divided into 2 parts. Section 4.1.1: Revised how "no spontaneous tropia" will be defined (ie just by control score) Section 4.1.6: Deleted sentence about imputed data being used, as the observed data will be used here. Section 4.1.7: Added section about Axial Length, which was added with Protocol Amendment II. Section 4.1.8: Added section about additional ocular biometric parameters, which were added with Protocol Amendment II. Section 4.2.1: Edited to state that 18-month subgroup analyses will be done only if there is an overall treatment group effect at 18 months. Redefined ranges for categories within baseline distance control subgroup and baseline refractive error subgroup. Edited how p-value will be calculated and interpreted. Table of summary of subgroups and expected directions of effect modification was added. Deleted paragraph about F-statistics. Included sentence about Forest plots that will be generated. Section 4.2.2: Deleted paragraph about analysis on baseline distance control subgroup of 3 to <5 points. Added paragraph about analysis on mean distance control at 18 months according to prescribed weaning status. Section 4.3.1: Added sentences about post-hoc risk ratio calculation of myopia progression for the overminus vs non-overminus group, both overall and in refractive error subgroups. Section 4.3.2: Corrected esotropia to esodeviation and SPCT to PACT. Originally Section 9.4: Deleted paragraph about Adverse Symptoms, as it was already being addressed in IXTQ section.

					Added sections on extension study.
2.1	V3.0 (Nov 22, 2019)	Amra Hercinovic		June 10, 2020	Revised per Leads' comments: Clarifications to study design and purpose; corrections to description of spectacle prescribing criteria.
2.2	V3.0 (Nov 22, 2019)	Danielle Chandler Amra Hercinovic		February 25, 2021	Updated analysis windows for 24- and 36-month visits. • Shortened the lower limit of the 24-month visit window from ≤18 months to ≤21 months. • Extended upper limit of 36-month visit analysis window from ≤36 months to ≤48 months Added sensitivity analysis of limiting primary analysis to visits completed within protocol window. Added sensitivity analysis of limiting primary analysis to participants who were not treatment crossovers. Added tabulation of outcomes by baseline SER subgroups to Section 5.5. Added calculation of correlation coefficients for biometric measures with 24/36mo SER to Section 5.3.

22

Amra Hercinovic
Digitally signed by Amra Hercinovic
Date: 2021.02.25 12:22:33 -05'00'

Amra Hercinovic, Statistician

Danielle Chandler
I am signing this document
2021-03-02 08:27-05:00

Danielle Chandler, Epidemiologist

Zhuokai Li
I am digitally signing this document
2021-02-25 13:58-05:00

Zhuokai Li, Senior Statistician

Angela Chen
Digitally signed by Angela Chen
DN: cn=Angela Chen, o, ou, email=angelachen@ketchum.edu, c=US
Date: 2021.03.04 18:26:14 -08'00'

Angela Chen, Protocol Co-Chair

Sergul Erzurum
Digitally signed by Sergul Erzurum
Date: 2021.03.07 22:33:11 -05'00'

S. Ayse Erzurum, Protocol Co-Chair

Susan Cotter
Digitally signed by Susan Cotter
Date: 2021.03.04 13:14:41 -08'00'

Susan Cotter,
PEDIG Network Co-Chair

Holmes, Jonathan - (jmholmes)
Digitally signed by Holmes, Jonathan - (jmholmes)
Date: 2021.02.25 10:15:17 -07'00'

Jonathan Holmes,
PEDIG Network Co-Chair

1. Design and Purpose of IXT5 Study

The objectives of the IXT5 study are to determine the long-term on-treatment effect of overminus treatment on distance IXT control score and the off-treatment effect of overminus treatment on distance IXT control score following weaning and 3 months off treatment. The participants will be randomly (1:1) assigned to the following treatment groups:

- Overminus Group (-2.50D over the cycloplegic refraction for 12 months, -1.25D over the cycloplegic refraction for 3 months, non-overminus correction for 3 months) Note that Protocol Amendment #2 (11/22/19) discontinued overminus lenses.
- Non-overminus Group [spectacles that fully corrected the astigmatism and anisometropia based on the cycloplegic refraction. The sphere power was based on SE in the least hyperopic eye as follows: 1) full correction for SE myopia, 2) sphere power that resulted in a plano SE lens (with symmetrical reduction in the most hyperopic eye, if needed).

Note: any mention of refractive error from this point on will refer to refractive error in the most myopic eye at baseline; this is because treatment was prescribed based on this eye and eligibility criteria, so this is the eye that will be used for any analyses involving refractive error.

2. Objective I: Effect of Overminus Lenses on Distance Control

2.1 Primary Analysis – Efficacy of Overminus Treatment (12 Months, On-Treatment)

The primary analysis will be a two-sided comparison of mean 12-month control of the distance exodeviation (mean of 3 measurements) between the treatment groups using an analysis of covariance (ANCOVA) model which adjusts for the following baseline factors to account for potential residual confounding: distance control, distance PACT, age, and refractive error (in the most myopic eye at baseline). The ANCOVA model will test the hypothesis that the treatment effect is different from zero (superiority hypothesis) using a type I error rate of 5%. The treatment group difference (overminus – non-overminus) and corresponding 95% confidence interval and p-value will be calculated. While use of ADHD medication was specified in the protocol as a baseline adjustment factor, this information was not collected at baseline due to an oversight. Therefore, this will not be one of the adjustment factors in the model.

2.1.1. Principles to be Followed in Primary Analysis

Model assumptions for the ANCOVA will be assessed, including linearity of the adjustment covariates (baseline distance control, distance PACT, age, and refractive error), normal distribution and equal variance across the treatment groups. The linearity assumption of the baseline covariates will be evaluated using descriptive scatterplots and by categorizing each of the baseline factors in the model to check for approximate linearity of the coefficients across ordered categories. A baseline covariate will be included as a continuous variable in the model if the assumptions for linearity are met for that covariate; otherwise, the baseline covariate will be categorized.

The 12-month distance control score for analysis for each participant is the mean of the 3 control assessments completed at that visit. When the protocol-specified three measures of control are not performed at the outcome exam, the mean of two measures will be used for analysis if only 2 distance control measures are completed; the single distance control score will be used for analysis if only 1 assessment of control is completed.

The primary analysis will follow a modified intention-to-treat principle, with all participants analyzed according to their randomized treatment group and with the following stipulations:

- Participants who are treatment crossovers (non-overminus group participants who are prescribed overminus; overminus group participants who have overminus spectacles formally discontinued before 12 months) will have their observed 12-month data analyzed provided they complete at least one distance control assessment at the 12-month outcome exam; otherwise, their mean distance control score will be imputed using multiple imputation.
- Participants who are prescribed IXT treatment other than overminus or non-overminus spectacles (e.g., surgery, vision therapy, patching) will have their mean distance control score imputed using multiple imputation, regardless of whether any control testing is completed at the 12-month visit.
- Participants who miss the 12-month visit or who do not complete any control testing at the 12-month visit, including treatment crossovers who miss or do not complete the 12-month visit, will also have their mean distance control score imputed using multiple imputation.
- To be included in the analysis, the 12-month visit must be completed no earlier than 270 days (≥9 months) and no later than 540 days (≤18 months) after randomization. If the 12-month visit is not completed within this analysis window, the distance control at 12-month visit will be imputed using multiple imputation.
- Data from participants randomized but found to be ineligible upon subsequent review of enrollment data will be included in the primary analysis.

The multiple imputation will be performed for missing data using Monte Carlo Markov Chain (MCMC) modeling. (See Section 2.4 below for further details.)

In April 2017, 28 participants were found to be incorrectly randomized because the wrong row from the randomization schedule was selected due to a programming error on the study website. (A list of these 28 participants is given in the Appendix of this analysis plan.) 20 of these 28 participants were randomized to the treatment group opposite the one they should have been randomized to if the correct row from the randomization schedule had been selected. 8 of the 28 participants were randomized to the correct treatment group by chance even though the incorrect record of the randomization schedule was chosen. As was approved by the DSMC, all 28 participants were continued in the (incorrectly) assigned treatment group (as recorded in tblPtRoster) and will be analyzed in that assigned treatment group.

2.1.2. Sensitivity Analysis

The following sensitivity analyses will be performed using the imputed dataset from the primary analysis unless otherwise specified.

2.1.3. Sensitivity Analysis #1: Complete Case Analysis (Excluding Out-of-Window Visits)

- The observed data (NOT the imputed dataset) will be used.
- All participants who complete one or more distance control assessments at a 12-month visit occurring between 9 to 18 months from randomization, including treatment crossovers and participants who are prescribed IXT treatment other than overminus or non-overminus spectacles, will have their observed 12-month data analyzed.
- Participants who miss the 12-month visit entirely or who do not complete any control testing at the 12-month visit will be excluded from the analysis.
- Participants who had the 12-month visit completed outside of the analysis window (<9 months or >18 months from randomization) will be excluded from the analysis.

2.1.4. Sensitivity Analysis #2: Complete Case Analysis (Including Out-of-Window Visits)

- The observed data (NOT the imputed dataset) will be used.
- Repeat sensitivity analysis in Section 2.2.1 above but *including* participants who had the 12-month visit completed outside of the analysis window (< 9 months or >18 months from randomization).

2.1.5. Sensitivity Analysis #3: Last Observation Carried Forward (LOCF) for Participants Starting Non-Study Treatment and Participants with Treatment Crossover

- Participants who are prescribed IXT treatment other than overminus or non-overminus spectacles for any reason at any time before the 12-month visit will have their mean distance control score from the last visit prior to starting non-study treatment for IXT carried forward as their mean distance control score at 12 months.
- Participants who switched from overminus to non-overminus treatment (or vice versa) before the 12-month visit will have their mean distance control score from the last visit prior to the treatment crossover carried forward as their mean distance control score at 12 months.
- Otherwise, data will be analyzed as in the primary analysis.

2.1.6. Sensitivity Analysis #4: Multiple Imputation for Participants with Treatment Crossover

- Participants who switched from overminus to non-overminus treatment (or vice versa) for any reason at any time before the 12-month visit will have their 12-month mean distance control score imputed using their mean control score from each of the follow-up visits before the treatment crossover occurred.
- Otherwise, data will be analyzed as in the primary analysis.

2.1.7. Sensitivity Analysis #5: Additional Adjustment Factors

- The primary analysis (a two-sided comparison of mean 12-month control of the distance exodeviation between the treatment groups using an ANCOVA model) will be repeated but will adjust for the following additional 4 factors: distance and near PACT and near stereoacuity. The model will adjust for these factors in addition to the initial baseline factors: distance control, distance PACT, age, and refractive error. The treatment group difference (overminus – non-overminus) and corresponding 95% confidence interval and p-value will be calculated.

2.1.8. Interim Analysis Summary

An interim monitoring plan has been developed to monitor for futility during the study. The interim monitoring plan is located at: F:\user\PEDIG\Studies\IXT\IXT5\Sample Size – Statistical Analysis\Statistical Analysis Plan\IXT5 SAP RW\Interim Monitoring\IXT5 Interim Monitoring Plan 11-10-17 (Final with Signatures).pdf

According to the IXT5 Interim Monitoring Plan, the DSMC may terminate the study early:

- If the treatment effect of overminus is not found to be statistically significant based on the partial 12-month data at the interim analysis stage, to be conducted when approximately 50% of the participants have completed 12 months of follow up.

OR

- If the treatment effect of overminus is not found to be statistically significant based on the analysis of complete 12-month data, to be conducted when all participants have completed 12 months of follow up (or have dropped from the study).

If the study is terminated early based on the analysis of partial or complete 12-month data, no further 15-month or 18-month visits will be required. The subsequent analyses involving 15-month and 18-month data outlined in this SAP will be descriptive only, with no statistical testing. The analyses involving 12-month data will be conducted using available 12-month outcome data in accordance with the specifications laid out in this SAP.

2.1.9. Multiple Imputation of Missing Values

The multiple imputation will be performed for missing data using Monte Carlo Markov Chain (MCMC) modeling, which includes the mean distance control scores at all follow-up visits prior to and including the 12-month visit, and the following baseline factors: distance PACT, age, and refractive error. The mean distance control score will be used in the MCMC model when multiple baseline measurements are available. The multiple imputation will be performed separately by treatment group to account for potential interaction between the treatment groups and the follow-up distance control scores. The number of imputations will be set to 100. For the participants who started IXT treatment other than overminus or non-overminus spectacles, only data from the visits prior to starting such treatment will be used in the MCMC model. Missing baseline data are not expected for distance PACT, age, and refractive error; however, if such missing data

occurs, it will be imputed in the MCMC model so that participants with such missing baseline data are included in the analysis.

If the study is not terminated early based on partial or complete 12-month data (as specified in Section 2.3), the missing mean distance control scores at all follow-up visits (including 15-month and 18-month visits) will be imputed again using an MCMC model that includes mean distance control scores at all follow-up visits prior to and including the 18-month visit, following the same principles specified above for the imputation of primary outcome at 12 months. Both the primary analysis for the 12-month outcome specified in Section 2 (including sensitivity analyses) and the secondary analysis for the 18-month outcome outlined in this SAP will be conducted using this newly imputed dataset and the results will be used in the final manuscript.

2.2 Secondary Analysis: Efficacy of Overminus After Treatment Discontinuation (18 Months, Off-Treatment)

The analyses specified in the secondary analysis will be performed on the full cohort which includes participants who were and were not prescribed full treatment between 0 to 12 months and weaning of overminus between 12 to 15 months. The full cohort analysis will be considered the main secondary analysis. In addition, these analyses will be repeated for the 18-month visit in a sensitivity analysis that will be limited to participants who were prescribed full treatment and weaning, defined as those who completed their 15-month visit before Protocol Amendment II discontinuing overminus treatment was approved by the Jaeb IRB on 12/18/19.. If retention of an on-treatment effect is stronger in participants who have full treatment/weaning vs. less than full treatment/weaning, the 18-month treatment effect could be diluted in the full cohort analysis vs. the limited cohort.

2.2.1 Treatment Group Comparison of Mean 18-month Distance Control

The analyses specified above in Section 2.1 will be repeated using the mean 18-month distance control score. As per the rationale in protocol change #2 of Protocol Amendment I, the 18-month comparison is a secondary analysis, hence no adjustment to alpha (5%) will be made to either the primary or the secondary analyses. It is acknowledged this could inflate the overall type I error rate, i.e. probability of one or more false positive findings, for the analyses of these outcomes.

The treatment group comparison of mean distance control at the 18-month visit will follow the same modified intention-to-treat principle as specified in Section 2.1.1. To be included in the analysis, the 18-month visit must be completed no earlier than 15 months and no later than 24 months after randomization. If the 18-month visit is not completed within the analysis window, the distance control at the 18-month visit will be imputed using multiple imputation (as specified in Section 2.1.9 above). Eighteen-month data for participants in the overminus group who do not discontinue overminus spectacles according to protocol will be included in analyses and will not be imputed, as will 18-month data from participants in the non-overminus group who start overminus.

2.3 Exploratory Analysis: Evaluate the Effect of Weaning

To evaluate the effect of weaning, the mean change in distance control score will be compared between the following visits within the overminus group using a paired t-test if the distribution is normal, or using a Wilcoxon signed rank test if the distribution is not normal:

- From 12-month visit to 15-month visit to assess change between full strength overminus treatment (-2.50 D) to half-strength (-1.25D) overminus treatment
- From 15-month visit to 18-month visit to assess 3 months of discontinuation of overminus treatment from half-strength (-1.25D) overminus treatment
- From 12-month visit to 18-month visit to assess a 6-month weaning strategy consisting of 3 months of half-strength (-1.25D) overminus treatment followed by 3 months discontinuation of overminus treatment

This analysis will be limited to participants who were prescribed weaning when the original protocol mandated it (i.e., prior to Protocol Amendment II which discontinued weaning). In addition, to be included in the analysis, the 15-month visit must be completed no earlier than 12 months and no later than 21 months after randomization. The imputed dataset specified in Section 2.1.9 will be used in all analyses outlined in Section 2.3.

Absolute treatment effect will be estimated by calculating retention of the 12-month on-treatment effect at 18 months (off-treatment). This will be done using an ANCOVA model that adjusts for the same covariates as the primary analysis (section 2.1), in addition to time, and the interaction between treatment group and time. A percentage estimating the treatment effect retained at 18 months will be reported from the model, as well as a 95% confidence interval.

3. Objective II: Effect of Overminus Treatment on Deterioration of IXT

3.1 Primary Analysis: Motor or Stereo Deterioration

Participants who meet either of the following at any visit will be considered to have met deterioration criteria.

- Motor deterioration: Control of the exodeviation measures 5 (constant exotropia) on all three assessments at distance and near. The exodeviation does not need to be constant throughout the entire exam provided that it is constant during all three assessments of control. (The study requires that a participant cannot be enrolled if the control of the exodeviation at near measures 5 on all three assessments at baseline; hence all participants in the study are eligible for motor deterioration.)
- Stereoacuity deterioration: Drop in near stereoacuity of at least 2 octaves (at least 0.6 log arcsec) from enrollment stereoacuity, or to nil, confirmed by a retest. (Participants with nil stereoacuity at enrollment will not be able to deteriorate with respect to a drop in near stereoacuity but are included in the analysis as they may have motor deterioration.)

Hazard of deterioration from a proportional hazards model will be compared by treatment group using the hazard ratio, and cumulative proportions deteriorating in each treatment group at 12 and 18 months will be reported. The cumulative proportion estimates from the proportional hazard model will be compared between treatment groups using a Z-test for proportions if at least one of the proportions is between 0.05 and 0.95. If both proportions are not between these two numbers, they will not be compared statistically. The proportional hazards model will adjust for baseline level of the components of the outcome, i.e. distance control, near control, and near stereoacuity. The population mean of the covariates in the proportional hazards model will be used for cumulative probability estimation. We will only adjust for these baseline-level components of the outcome because deterioration rates were low in IXT2,[1] making adjustment for more than a few covariates potentially problematic. The observed data will be used in these analyses.

Any observation with a missing value for any of the baseline factors will be excluded from the analysis. However, the number of missing values is expected to be low given that the testing of these baseline factors is required at enrollment by protocol. If either motor or stereo outcome data are missing at a follow-up visit and these data are available at a later visit, the deterioration status will be carried forward from the visit previous to the missing visit. The participants who are lost to follow up without meeting deterioration criteria will be considered censored as of the last completed follow-up visit.

The participants who start non-study treatment (i.e., treatment other than overminus or non-overminus) for IXT will be censored 0.1 month after the visit when the non-study treatment was prescribed. Participants who are treatment crossovers (non-overminus group participants who are prescribed overminus; overminus group participants who have overminus spectacles formally discontinued against protocol) will have observed deterioration status data included in analysis, i.e., will not be censored on the basis of starting a non-randomized treatment (i.e., any treatment other than the one assigned by the study at randomization). The proportional hazards model time variable will be discrete with times corresponding to the follow-up time of study visits.

The proportional hazards assumption will be checked by testing for a time by treatment interaction in the model, as well as by visual inspection of Kaplan-Meier time-to-event curves. In the event the proportional hazards assumption is not met, an alternative analysis such as a weighted, stratified comparison of Kaplan-Meier estimates will be considered.

3.2 Secondary Analysis: Cause-Specific Deterioration

If the overall deterioration treatment group comparison (Section 3.1) is performed and is significant, then the cumulative incidence of cause-specific deterioration (stereo ONLY and motor ONLY) by 12-months and 18-months will be estimated and compared between the treatment groups using the cumulative incidence function to account for the competing risks of starting treatment due to deterioration from a cause other than the

cause of interest, or starting treatment against protocol. To control for multiplicity, alpha will be split between the two time points (12 and 18 months), and a gatekeeping strategy will be used within each time point, i.e., the 18-month difference will be tested only if the 12-month difference is statistically significant. These analyses are considered post-hoc, as they were not included in version 1.0 of the protocol or SAP. Because using stereoacuity as the (sole) outcome for IXT has ample precedent in the literature, an additional analysis will be conducted using a post-hoc alternative definition of deterioration which classifies the participant's condition as deteriorated only if the participant met stereo deterioration criteria. *Unlike for the analysis of overall deterioration, participants with nil stereo at baseline will be excluded from this analysis.* The analysis will be adjusted for baseline near stereoacuity.

4. Objective III: Other Clinical Outcomes, Exploratory Analyses, and Safety

The analyses specified below will use the observed data and only the visits that occurred within the analysis window will be included. For all outcomes, the data will be tabulated and means and standard deviations, or proportions, will be calculated for each of the treatment groups. The type I error rate for analyses of the secondary outcomes specified in the sections below will be controlled using false discovery rate (FDR) to account for the multiple outcomes. The outcomes will be divided into 2 groups for this purpose, with FDR controlled at the 5% probability level within each group. The first group is comprised of the clinical measures (no spontaneous tropia, near control, angle magnitude, stereoacuity, axial length and other biometric measures, and refractive error) and the second is comprised of the symptom and quality of life measures.

4.1 Other Clinical Outcomes at 12 and 18 Months

For the 18-month visit, all analyses will be repeated, limited to participants who were prescribed weaning (i.e., prior to weaning being discontinued per Protocol Amendment II). This will be done because if retention of an on-treatment effect is stronger in participants who have full weaning vs. less than full weaning, the 18-month treatment effect could be diluted in the full cohort analysis vs. the limited cohort.

4.1.1 No Spontaneous Tropia

The proportion of participants with no spontaneous tropia will be compared between the two treatment groups using a two-sided Barnard's test (with alpha of 0.05) at 12 months and 18 months. The difference in proportions, a two-sided 95% confidence interval, and p-value on the difference in proportions will be calculated.

No spontaneous tropia during control testing at the 12-month (and 18-month) exams means the following must have been true during the examination:

- Score of ≤ 2 (2 or better) on all three assessments of control at distance and at near

While the definition originally specified "No spontaneous tropia at any time during the exam at distance or near," the way the question was posed on the CRF did not allow us to differentiate between spontaneous and non-spontaneous tropia for the period of time

outside of control testing for patients who were classified as having intermittent tropia. Therefore, we changed the definition to only be based on presence of spontaneous tropia during control test scores, as detailed above.

The proportion of participants with no spontaneous tropia during control testing at 12 months (and 18-months) will also be tabulated for each treatment group within the subgroups specified below, if the overall effect was significant. The estimate or treatment difference and 95% confidence interval will be reported.

- Spontaneous tropia (as defined above) at baseline (Yes vs. No)
- Severity of baseline distance control (2 to <3, 3 to <4, 4 to 5)

4.1.2 Change in Distance Control

Distance control will be reported as the distributions of baseline level of control, 12-month (primary outcome), and 18-month control, and *change in control* from baseline to 12 months, and 18-months, including the proportion with ≥ 1-point change in control and proportion with ≥ 2-point change.

The proportion of participants with ≥ 1-point improvement in distance control between baseline and 12 months (18 months) will be compared between the treatment groups using a two-sided Barnard's test with alpha of 0.05, with calculation of a two-sided 95% confidence interval and p value on the difference in proportions. The proportion of participants with ≥2-point improvement in distance control between baseline and 12 months (18 months) will be compared between the treatment groups similarly.

4.1.3 Additional Sensitivity Analyses

The treatment group comparisons of distance control score at the 12-month and 18-month visits will be repeated excluding the 28 participants whose treatment assignment was based on the incorrect row from the randomization schedule (Section 2.1.1), regardless of whether the incorrectly-assigned value matched the assignment that would have been chosen if the correct record from the randomization schedule was used. The imputed dataset specified in Section 2.1.9 will be used. The results will be compared to the original analyses to assess if bias was introduced by including these 28 participants who were included in in the primary analyses despite being incorrectly randomized.

4.1.4 Near Control

At the 12 and 18-month time points, near control will be evaluated similarly to the distance control primary analysis (Section 2).

4.1.5 Angle Magnitude by PACT

At the 12- and 18-month time points, a two-sided treatment group comparison of magnitude of the deviation by Prism Alternate Cover Test (PACT) will be performed using an ANCOVA model which adjusts for baseline PACT. The treatment group difference, 95% confidence interval and p value will be calculated. The analysis will be completed separately for distance PACT and at near PACT.

4.1.6 Stereoacuity

At the 12 and 18-month time points, a two-sided treatment group comparison of near stereoacuity (measured with the Preschool Randot Test) will be performed using an ANCOVA model which adjusts for baseline stereoacuity. The treatment group difference, 95% confidence interval, and p value will be calculated.

4.1.7 Axial Length Measurement

At the 18-month time point, mean axial length measurement will be compared between treatment groups using an ANOVA model. The treatment group difference and a 95% confidence interval will be calculated. Adjustment for baseline is not possible given that these measurements were added late in the study (with Protocol Amendment II) and therefore not collected at baseline.

4.1.8 Additional Ocular Biometric Parameters

At the 18-month time point, mean flat corneal radius, anterior chamber depth, and lens thickness will each be compared (if available) between treatment groups using an ANOVA model. The treatment group difference and a 95% confidence interval will be calculated. Adjustment for baseline is not possible given that these measurements were added late in the study (with Protocol Amendment II) and therefore not collected at baseline.

4.1.9 Refractive Error

At the 12- and 18-month time points, a two-sided treatment group comparison of spherical equivalent refractive error will be performed using an ANCOVA model which adjusts for baseline refractive error and baseline age. The latter was added as a post-hoc adjustment factor, as age is expected to be associated with refractive error at the 12-month (and 18-month) time points. The treatment group difference, 95% confidence interval. and p-value will be calculated. This analysis will be done combining refractive error obtained by subjective refraction or retinoscopy and separately from refractive error obtained by autorefraction.

4.1.10 Quality of Life

At both the 12-month and 18-month time points, a two-sided treatment group comparison of the child, and each of the three parent scales in the IXTQ HRQOL (Quality of Life) will be performed each using the Wilcoxon Rank Sum Test as it is expected that scores will not be normally distributed and will have a substantial ceiling effect. The treatment group difference, 95% confidence interval, and p value will be calculated for both the Rasch-scored data[1] and for 0 to100-scaled data. The reported p-value will be derived from the analysis of Rasch-scaled data, while to aid in interpretation, the 0 to 100-scaled data will be used when reporting the median values for each treatment group.

[1] Leske DA, Holmes JM, Melia M, on behalf of Pediatric Eye Disease Investigator Group. Evaluation of the Intermittent Exotropia Questionnaire using Rasch analysis. JAMA Ophthalmol 2015;133:461-5.
Leske DA, Hatt SR, Liebermann L, Holmes JM. Evaluation of the adult strabismus-20 (as-20) questionnaire using rasch analysis. Invest Ophthalmol Vis Sci 2012.
Leske DA, Hatt SR, Liebermann L, Holmes JM. Lookup tables versus stacked rasch analysis in comparing pre- and postintervention adult strabismus-20 data. Transl Vis Sci Technol 2016;5:11.

A potential secondary manuscript, to be detailed in a separate SAP, will describe quality of life, IXT symptoms, and spectacle wear issues, the relationships among these, and relationships between these and clinical measures.

4.1.11 IXT Symptom Survey

A 7-item survey of IXT-related symptoms will be completed by the child at the enrollment, 6-month, 12-month, and 18-month outcome exams. For each IXT-specific symptom, the child is asked how often it occurs, using the response choices of "never" "sometimes", and "all the time." The distribution for each individual item will be tabulated for each treatment group for each outcome time point.

A 7-item survey consisting of IXT-related symptoms and problems with spectacle wear will be completed by parents of participants at the enrollment, 6-month, 12-month, and 18-month exams. The parent is asked how often symptoms or problems occur, using the response choices of "never," "almost never," "sometimes", "often", "almost always." The distribution for each individual item will be tabulated for each treatment group for each outcome time point.

4.1.12 Compliance with Spectacle Wear

Compliance with spectacle wear will be assessed at the 6-month, 12-month, 15-month, and 18-month exams. Parents will provide an estimate of the proportion of the time their children wore their spectacles. Proportion of time worn will be described as excellent (76% to 100%), good (51% to 75%), fair (26% to 50%), or poor (≤ 25%). The distribution of compliance will be tabulated for each treatment group for each time point.

4.2 Exploratory Analyses

The following exploratory analyses will use observed or imputed data (see LOCF, below), as follows:

- All participants who complete one or more distance control assessments at a specific visit within the corresponding analysis window, including treatment crossovers, will have their observed data analyzed, except for participants who are prescribed non-study treatment (i.e., IXT treatment other than overminus or non-overminus refractive correction), who will have the distance control from the (first) visit when non-study treatment was prescribed carried forward.
- Participants who miss the study visit entirely or who do not complete any control assessments at the visit will be excluded from the analysis for that visit.
- Participants who had the visit completed outside of the analysis window will be excluded from the analysis for the visit.

4.2.1 Mean Distance Control in Baseline Subgroups

The treatment group comparisons of 12-month distance control will be assessed in subgroups based on baseline factors. Subgroup comparisons for 18-month distance control will be explored only if there is an overall treatment group difference in distance control at 18 months. All planned subgroup analyses will repeat the primary analysis (Section 2, excluding the sensitivity analysis). The specific subgroups of interest are:

- Baseline distance control score by severity (2 to <3, 3 to <4, 4 to <5 points)

- Baseline age group (3 to <7 years vs. 7 to <11 years)
- Spherical equivalent refractive error level (-6.00 to -0.50D vs. -0.375 to +1.00D)
- Accommodative convergence/accommodation (AC/A) ratio (<2.5, 2.5 to 6.0, >6)

Exploratory analyses performed on the IXT3 pilot study data suggested that overminus spectacles might be effective for participants with very poor baseline distance control (4-<5 points), a finding that was somewhat unexpected. We would like to further explore whether treatment effect varies by baseline control. Because children 7 to < 11 years of age likely have greater accommodative demands (e.g., schoolwork) than children 3 to < 7 years of age, and because the visual systems of the two groups are at different stages developmentally, it is possible that their response to overminus lenses might be different. Since participants with hyperopia and participants with myopia may use their accommodative system differently, the response to overminus lenses may be different. We might expect that participants with high AC/A ratios would respond to overminus spectacles better than participants with lower AC/A ratios given that one of the mechanisms of action of overminus is to stimulate accommodative convergence, thereby reducing the angle of exodeviation and allowing fusion. The IXT3 pilot study showed a similar treatment effect of overminus in participants with AC/A <2.5 and >=2.5 but had very few participants with very high AC/A (i.e., larger than 6).

Subgroup analyses will also be conducted for gender and race/ethnicity in accordance with NIH guidelines; however, no effect modification is expected for these factors.

A summary of the subgroups and expected directions of effect modification is shown in Table 1.

Table 1 – Subgroup Analyses

Factor	Grouping	Expected Direction	Rationale
Baseline age (years)	3 to <7 7 to <11	Better treatment effect expected in 3 to <7 group	Younger children typically have a greater accommodative amplitude and engage in less sustained near activities than older children and may more completely accommodate, thereby inducing accommodative convergence which is

			a postulated mechanism for overminus
Baseline distance control	2 to <3 3 to <4 4 to <5	Larger treatment effect expected in participants with worse control (4 to <5)	Observed in IXT3. Possibly, participants with worse control have more room to improve / less ceiling effect. They also are expected to have more regression to the mean.
Refractive error (D)	-6.00 to -0.50 (myopic) -0.375 to +1.00 (not myopic)	Better treatment effect expected in non-myopes	Myopes are likely to get more myopic during the study, which would lessen the true effect of overminus glasses.
AC/A ratio	<2.5 2.5 to 6.0 >6	Better treatment effect expected in participants with higher AC/A ratios (>2.5)	Higher AC/A is associated with greater convergence per unit of accommodation, and accommodative convergence is a postulated mechanism of overminus effect

The descriptive data tabulations and figures of the subgroups will be based on the observed data. The subgroup analysis will not be done if the sample size in a subgroup is less than 20 participants for any subgroup level within treatment group.

A Forrest plot will be generated, which will display estimates of difference between treatment groups and 95% confidence intervals for every subgroup. P-values will not be reported.

In addition to the baseline factors specified above, the treatment group comparisons of 12-month and 18-month (if there is an overall effect) distance control will also be assessed in subgroups based on the use of ADHD medications at any time between baseline and the 12-month visit (yes/no). ADHD medication may reduce accommodation and could make overminus treatment less effective. Note: As the criteria used to categorize use of ADHD medications is based on post-randomization factors (i.e., use of medication during the follow-up), this could potentially introduce bias into the treatment

group comparison. The treatment effect in these subgroups will be interpreted with particular caution.

4.2.2 Mean Distance Control at 18 Months According to Prescribed Weaning Status

Of the 355 expected 18-month visits (as of 11/5/19), approximately 283 participants (80%) will have completed a full 3-months of prescribed weaning before discontinuing overminus (or non-overminus); approximately 35 to 54 (20% to 25%) participants will have completed partial weaning, and approximately 18 to 37 participants (5% to 10%) are expected to have no weaning at all per Protocol Amendment II. The summary statistics for distribution of 18-month distance control and change from baseline in distance control will be tabulated according to treatment group and whether full, partial, or no prescribed weaning.

4.2.3 Participants with Baseline Control of 3 to 5 Points (Spontaneous Tropia)

Participants with baseline distance control of 3 to 5 points may have a better likelihood of improving ≥2 points. The mean distance control of these participants is indicative of the presence of spontaneous tropia on at least one of the 3 control measures, while participants with mean distance control scores of 2 to <3 points had a phoria on one or more of the control measures.

If there is no evidence of a treatment effect in the subgroup of participants with mean distance control scores at baseline of 2 to <3 points (from Section 4.2.1), the distance control primary analysis (Section 2, excluding sensitivity analysis) and secondary analysis (Section 2.3) will be repeated in exploratory analyses limited to the cohort of participants with baseline distance control of 3 to 5 points. This will be done at both the 12-month and 18-month time points.

[Note: this is included in the SAP only because it appeared in the Protocol. It is recognized that in the event this analysis is carried out as specified, strict control over the type I error rate will have been lost.]

4.3 Safety Analyses

4.3.1 Refractive Error at 12 and 18 Months

The proportion of participants with >1 D increase in myopia or in hyperopia from baseline to 12 months and from baseline to 18 months will be tabulated separately by direction of change (myopic or hyperopic), and treatment group, for refractive error as measured by subjective refraction or retinoscopy and also for refractive error as measured by autorefraction. The risk ratio for myopia progression of >1 D will be calculated for the overminus group vs the non-overminus group. This will be done both overall, as well as stratified by baseline refractive error groups as follows: myopic (-0.50D to -6.00D), emmetropic (-0.375D to +0.375D), and hyperopic (+0.50D to +1.0D). [This latter is a post-hoc analysis added to help quantify risk after an increased risk of myopia was identified in the overminus group during a DSMC review of safety and efficacy data. The

analysis stratified by baseline refractive error will only be done for refractive error measured by retinoscopy or subjective refraction.]

4.3.2 Development of Esodeviation

Frequency and percent developing any esodeviation will be tabulated by treatment group, indicating the magnitude of the esodeviation (by PACT) and whether it was a constant tropia, intermittent tropia, or a phoria.

4.3.3 Reduction of Distance Visual Acuity

Any cases of reduced visual acuity in best refractive correction of ≥2 logMAR lines in either eye will be tabulated by treatment group.

4.4 Protocol Adherence and Additional Tabulations

The following tabulations will be performed:

- A flow chart accounting for all participants according to treatment group for all visits.
- Visit completion rates for each follow-up visit according to treatment group.
- Protocol deviations according to treatment group.
- Baseline demographic and clinical characteristics according to treatment group, with summary measures such as mean and standard deviation, when applicable
- Number of and reasons for unscheduled visits and phone calls
- Number of and reasons for treatment crossovers according to randomized treatment group
- Number of and reasons for non-study treatment according to randomized treatment group

Appendix – List of Incorrectly Randomized Participants

Due to an error in web programming that was caught in late April 2017, 28 participants were randomized incorrectly – the web program selected the wrong row of the randomization schedule.

Randomized to Incorrect Treatment (N=20)

The following 20 participants were randomized to the treatment group opposite the one they should have been randomized to if the web had selected the correct row of the randomization schedule.

1. I05-008-0035
2. I05-008-0037
3. I05-079-0337
4. I05-079-0336
5. I05-094-0087
6. I05-180-0269
7. I05-180-0272
8. I05-180-0275
9. I05-273-0098
10. I05-307-0049
11. I05-307-0050
12. I05-307-0054
13. I05-307-0053
14. I05-319-0011
15. I05-319-0007
16. I05-319-0008
17. I05-328-0019
18. I05-328-0020
19. I05-328-0017
20. I05-331-0004

Randomized to Correct Treatment (by chance) (N=8)

The following 8 participants were randomized to the correct treatment group (by chance) even though the wrong record of the randomization schedule was chosen.

1. I05-008-0036
2. I05-033-0272
3. I05-079-0345
4. I05-079-0342
5. I05-180-0273
6. I05-328-0021
7. I05-331-0006
8. I05-331-0007

The above lists were provided by R. Kraker at Jaeb Center for Health Research.

INTERMITTENT EXOTROPIA STUDY 5 (IXT5)
Extension Study

A Randomized Clinical Trial of Overminus Spectacle Therapy for Intermittent Exotropia –

Extended Through 36 Months

5. IXT5 Extension Study

5.1. Design and Purpose of IXT5 Extension Study

The objective of the IXT5 extension study is to compare long-term change in refractive error between participants originally randomized to overminus spectacles or non-overminus spectacles as part of the 18-month randomized trial.

5.2. Primary Analysis – Comparison of Refractive Error Between Treatment Groups at 24 Months and 36 Months

The primary analysis will be a two-sided comparison between treatment groups of change from baseline in spherical equivalent refractive error in the most myopic eye at baseline at 24 months and at 36 months. The comparison(s) will be made using an analysis of covariance (ANCOVA) model that adjusts for baseline spherical equivalent refractive error to improve power and account for potential residual confounding. The comparison will test the null hypothesis that the difference in change in refractive error between the two groups is zero using a type I error rate of 5%. The treatment group difference (overminus – non-overminus) and corresponding 95% confidence interval and p-value will be calculated. These analyses are essentially safety analyses; thus, there will be no adjustment to type I error for multiplicity.

5.2.1. Principles to be Followed in Primary Analysis

Model assumptions for the ANCOVA will be checked as described in Section 2.1.1, including linearity of baseline refractive error, normal distribution, and equal variance across treatment groups. If baseline refractive error was included as a continuous variable in the analysis from Section 2.1, it will be included as a continuous variable in this model as well; otherwise, it will be categorized.

There will be no imputation for missing data; however, baseline characteristics of participants with and without follow up will be tabulated at each time point. In the event there is a clinically relevant imbalance between those with and without follow up, a propensity score-weighted analysis will be performed as a sensitivity analysis.

The primary analysis will follow a modified intention-to-treat principle, with all participants analyzed according to their randomized treatment group. To be included in the analysis, the 24-month visit must be completed no earlier than 630 days (≥21 months) and no later than 899 days (<30 months) after randomization. If the 24-month visit is not completed within this analysis window, the 18-month visit data will be used instead (if available and if within 24-month window). The 36-month visit must be completed no earlier than 900 days (≥30 months) and no later than 1440 days (≤48 months after randomization.

As a sensitivity analysis, the primary analysis will be repeated but limited to participants who completed each visit within the protocol window of +/- 3 months.
A separate sensitivity analysis will repeat the primary analysis but be limited to participants who were not treatment crossovers at any point in the study.

5.3. Secondary Analysis—Comparison of Axial Length and Other Biometric Measurements Between Treatment Groups at 24 and 36 Months

At the 24-month and 36-month time points, mean axial length measurement, flat corneal radius, anterior chamber depth, and lens thickness (if available) will each be compared between treatment groups using an ANOVA model. Adjustment for baseline is not possible given that these measurements were added late in the study (with Protocol Amendment II) and therefore not collected at baseline. Treatment group differences, 95% confidence intervals, and p-values will be calculated. The type I error rate in this secondary analysis will be controlled using false discovery rate (FDR) to control the FDR at 5% and to adjust the p-values for multiplicity.

Longer axial length is expected to be related to more myopic progression. The other biometric measures may or may not be related to myopic progression. To assess this, Spearman correlation coefficients will be calculated for the relationship between the biometric measures (mean axial length measurement, flat corneal radius, anterior chamber depth, and lens thickness) and refractive error at the 24-month and 36-month time points. Correlation coefficients with an absolute value of 0.60 or higher will be considered strong correlations.

5.4. Exploratory Analyses

The analyses described in the sections below will be exploratory because treatment for IXT was at investigator discretion after 18 months. For each outcome, a difference between treatment groups and 95% CI will be estimated, but p-values will not be calculated. If the outcome is continuous, the difference and 95% CI will be calculated using an ANCOVA model adjusted for the baseline value of the measure, as was done for the continuous outcomes in Section 4. If the outcome is categorical (i.e., a proportion), a two-sided Barnard's exact test will be used as was done in Section 4.1.1.

5.4.1. Refractive Error Proportions at 24 and 36 Months

The proportion of participants with >1 D, ≥ 2 D, and ≥ 3 D increase in spherical equivalent refractive error (in the myopic direction) from baseline to 24 months and baseline to 36 months will be tabulated by treatment group.

5.4.2. Distance Control

Mean distance control will be calculated and reported by treatment group as the distribution of baseline control, 24-month control, 36-month control, and change in control from baseline to 24- and 36-months.

5.4.3. Near Control

Near control will be reported by treatment group at 24-months and 36-months similarly to the distance control in Section 5.4.2.

5.4.4. No Spontaneous Tropia

The proportion of participants with no spontaneous tropia (defined in section 4.1.1) will be reported for each treatment group at 24-months and 36-months.

5.4.5. Angle Magnitude by PACT

At the 24- and 36-month time points, the distribution of the magnitude of the deviation by Prism Alternate Cover Test (PACT) will be tabulated by treatment group. This will be completed separately at distance and at near.

5.4.6. Stereoacuity

At the 24- and 36-month time points, the distribution of the near stereoacuity by testing with the Preschool Randot Test will be tabulated by treatment group.

5.4.10 Cycloplegic Autorefraction at 24 and 36 Months

The change from baseline in cycloplegic autorefraction will be tabulated for each treatment group and descriptive statistics will be calculated at 24- and 36-months for participants who have these data (autorefractors are not available at all sites). The magnitude of refractive error (as determined by autorefraction) will be tabulated by treatment group and descriptive statistics will be calculated.

5.5 Additional Tabulations

5.5.1 Development of Esodeviation

The number and percent of participants with an esodeviation at 24 months and 36 months will be tabulated by treatment group. The magnitude of the esodeviation (by PACT) and whether it was a constant tropia, intermittent tropia, or a phoria will be specified.

5.5.2 Reduction of Distance Visual Acuity

Any cases of reduced visual acuity of ≥2 logMAR lines from baseline in either eye when wearing best refractive correction will be tabulated by treatment group.

5.5.3 Additional Treatment

Any additional treatment used after 18 months was at investigator discretion and will be reported by treatment group. At the 24- and 36-month visits, treatments used since the last visit will be tabulated according to treatment group.

5.5.4 Tabulations by Baseline Refractive Error Subgroups

Summary statistics for refractive error at 24 and 36 months will be tabulated by baseline refractive error subgroups for each treatment group. Baseline refractive error groups will be defined as follows: myopic (-0.50D to -6.00D SE), emmetropic (-0.375D to +0.375D SE), and hyperopic (+0.50D to +1.00D SE). Additionally, biometric measures (Section 5.3) will also be tabulated by these baseline subgroups, as will any outcomes in Section 5.4 that are determined to be of further interest.

Reference

1. Pediatric Eye Disease Investigator Group. Three-year observation of children 3 to 10 years of age with untreated intermittent exotropia. Ophthalmology 2019;126:1249-60.