Intermittent Exotropia Study 7

(IXT7)

Randomized Trial of Full-Time Occlusion Therapy for Intermittent Exotropia in Children

STATISTICAL ANALYSIS PLAN

Revision History

Version Number		Author	Approver	Effective Date	Revision Description (Including Sections Revised)
SAP	Protocol				
1.0	V1.1 17Oct2022	A. Hercinovic / M. Melia	R. Kraker	18Oct2022	Initial version (Completed prior to study initiation)
1.1	V1.1 17Oct2022	R. Kraker	M. Melia	10Jan2023	Corrected typographical errors with respect to range of change from baseline in control (-5 to 2.7 instead of -5 to 3.0) and removed diplopia by parent and/or child yes/no outcomes from correlations with compliance data since are not continuous outcomes.
1.2	V1.2 26Jul2024	R. Kraker Y. Zhu	W. Beaulieu	06May2025	Clarified that mean distance control score at baseline must be 2.0 or greater, not 2.3 or greater (Section 2). Clarified criteria for sensitivity analysis excluding participants with interocular difference of 2 lines (Section 2). Clarified that the false discovery rate will be controlled for secondary outcomes using the two-stage step-up procedure (Section 3). Edited threshold for exploratory temperature sensor adherence to match data published after SAP drafted (Section 5). Required subgroup analyses to have at least 10 observations per treatment arm for analysis (Section 7). Updated diplopia tabulations to match case report form (Section 8). Changes made after reviewing outcome data.

Signature Page

Signature of Author:

Yufeng Zhu
I am digitally signing this document
2025-05-06 11:29-04:00

Ray Kraker
I agree to the terms defined by the placement of my signature in this document
2025-06-02 08:21-04:00

Signature of Approver:

Wesley Beaulieu
I agree to the terms defined by the placement of my signature in this document
2025-05-06 11:22-04:00

1 Purpose of IXT7 Study

The objective of the IXT5 study is to determine whether 3 months of full-time patching is more effective than observation for improving distance control of IXT (on-treatment outcome). A total of 72 participants will be randomly (1:1) assigned to the following treatment groups:

- Observation for 3 months – no treatment other than refractive correction
- Full-time (FT) patching (all waking hours) for 3 months (right eye on even days, left eye on odd days), in addition to refractive correction

2 Primary Analysis – Efficacy of Full-Time Patching (3 Months, On-Treatment)

The primary outcome is change from baseline to 3 months in mean (of 3 measurements) distance control score (mean 3-month score – mean baseline score). Single control scores are integer-valued and range from 0 to 5 with 0 indicating no deviation (no tropia or phoria) and 5 indicating constant tropia. To meet eligibility criteria, the mean score must be $\geq$2.0 and $\leq$5.0 points. Hence, the change in mean control score is a quasi-continuous, bounded measurement with possible values ranging from -5.0 to 3.0 points, with negative values indicating improvement and positive values indicating worsening.

The primary analysis will be a two-sided comparison of the mean change in 3-month control of the distance exodeviation between the treatment groups using an analysis of covariance (ANCOVA) model which adjusts for the baseline distance control and distance PACT (prism and alternate cover test) angle. 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%. An adjusted treatment group difference (FT patching – Observation) and corresponding 95% confidence interval and p-value will be estimated from the ANCOVA model.

2.1 Principles to be Followed in Primary Analysis

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

When three measures of distance control are not performed at the outcome exam, the mean of two measurements, or failing that, a single measurement, will be used if available.

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

- For participants who receive non-randomized treatment between baseline and 3 months, the distance control score data from the early 3-month visit will be used for the primary outcome. Early outcome visits are not required to fall within the 3-month visit window to be included in analyses.

- Participants who miss the 3-month visit or do not complete any control testing at the 3-month visit, and who do not have an early 3-month outcome visit, will have their mean distance control score imputed using multiple imputation.

- For all participants other than above, to be included in the analysis, the 3-month visit must be completed within the analysis window defined as 46 to 137 days from randomization; otherwise, the distance control at the 3-month visit will be imputed using multiple imputation.

The multiple imputation for missing data will be performed using Monte Carlo Markov Chain (MCMC) model. (See Section 2.3 below for further details.)

2.2 Sensitivity Analyses

Sensitivity analyses of the primary outcome will repeat the primary analysis with the following modifications:

2.2.1 Sensitivity Analysis #1: Tipping Point Analysis to Test the Missing at Random Assumption of the Primary Analysis

MCMC multiple imputation assumes that missing data are missing at random (MAR). To test this assumption, a tipping point analysis will be performed. In IXT7, the MAR assumptions means that whether the 3-month control score is missing or observed may be a function of the baseline PACT angle and baseline control score, which are included in the imputation and analysis models, but are not a function of the missing data being imputed, after conditioning on the baseline control and PACT angle. This assumption cannot be tested directly as missing data are unknown. The tipping point analysis will adjust the imputed values using a shift parameter to determine how severe the departure from MAR must be to change the conclusion of the analysis with respect to rejecting or failing to reject the null hypothesis.

If the 3-month outcome in one group is significantly better than the other (significantly lower mean control score as determined in the primary analysis), a shift parameter will be applied to the imputed values in that group to determine the tipping point at which the results of the primary analysis are nullified. That is, if one group is found to be superior ($P \leq .05$), the tipping point will identify the shift parameter necessary to yield $P > .05$. Conversely, if the null hypothesis is not rejected ($P > .05$), two tipping points will be identified – one that would make patching group superior and one that would make observation group superior. In either case, this tipping point will be evaluated to determine if it is plausible, or even possible given the control score range is bounded. If not possible, the MAR assumption does not affect results, regardless of plausibility. If possible but not plausible, the MAR assumption is likely reasonable. If plausible, the sensitivity of study results to missing data will be reported as a potential study weakness.

2.2.2 Sensitivity Analysis #2: Complete Case Analysis (Excluding Imputed Control Scores)

Participants whose 3-month control score was imputed in the primary analysis will be excluded from the analysis.

2.2.3 Sensitivity Analysis #3: Complete Case Analysis (Excluding Imputed Control Scores and Early Outcome Visits)

Repeat the sensitivity analysis in section 2.2.2 excluding data from participants who completed an early outcome visit.

2.2.4 Sensitivity Analysis #4: ITT Excluding Participants with IOD 2 LogMAR Lines at Enrollment

The primary analysis will be repeated excluding participants of any age with IOD of 2 logMAR lines (8 or more letters for E-ETDRS and 2 logMAR lines for ATS-HOTV).

2.3 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 score at baseline and the change in mean distance control score from baseline to the 3-month (or early outcome) visit, and the distance PACT at baseline. Imputed values will be limited to the range of possible changes in mean control score (-5.0 to 3.0), using options available in the imputation software. If necessary, observed 3-month control scores may be transformed to achieve (or approximate) the normal distribution prior to imputing, and then back-transformed for analysis.

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. Missing baseline data are not expected for distance PACT angle; however, if such missing data occurs, it will be imputed in the MCMC model so that participants with missing baseline data are included in the analysis.

3 Secondary Analyses

All secondary analyses described below will use observed data only without imputation for missing data. Only data from 3-month outcome visits that occurred within the analysis window (or data from early 3-month outcome visits) will be included in analysis. For all secondary outcomes, the data will be tabulated, and means and standard deviations, or numbers with the outcome and percentages, will be calculated for each of the treatment groups. The false discovery rate will be controlled at 5% using the two-stage step-up procedure[1] to account for multiple secondary outcomes. Reported p-values and confidence intervals will incorporate the FDR adjustment.

3.1 Near Control

At 3 months, change in near control (mean of 3 measurements) will be evaluated similarly to distance control in the primary analysis using an ANCOVA model that adjusts for the mean near control score at baseline and the near angle by PACT at baseline. If only 2, or 1, near control measurements are available, the 3-month near control score will be calculated based on the available measurements, as will the change from baseline.

Additionally, the proportion of participants with 2 or more points improvement from baseline in mean near control will be tabulated and compared between the treatment groups at 3 months using a two-sided Barnard's test, separately for distance and near. The treatment group difference and 95% confidence interval will be estimated by inverting Barnard's test, using the method of Agresti and Min.[2]

3.2 Angle of Deviation

At 3 months, the angle of deviation by PACT will be described using mean, SD, and range, and compared between treatment groups using an ANCOVA model, adjusting for baseline PACT angle. The analysis will be completed separately for the distance and near angles. Esodeviations will be represented using a negative sign to differentiate them from exodeviation.

Additionally, the proportions of participants with improvement of distance angle and improvement of near angle (separately) at 3 months greater than the test/retest variability will be compared between the treatment groups using Barnard's test. The treatment group difference and 95% confidence interval will be estimated by inverting Barnard's test, using the method of Agresti and Min.[2] The published test-retest criteria[3] are as follows:

At distance:

- If deviation magnitude $\leq$20 Δ, $\geq$4 Δ decrease exceeds test-retest variability
- If deviation magnitude >20 Δ, $\geq$8 Δ decrease exceeds test-retest variability

At near:

- If deviation magnitude $\leq$20 Δ, $\geq$7 Δ decrease exceeds test-retest variability
- If deviation magnitude >20 Δ, $\geq$13 Δ decrease exceeds test-retest variability

Development of an esodeviation $\geq$4 Δ will not be counted as an "improvement in angle" in these analyses. Also, although the same guidelines above apply to classifying increases in angle as beyond test-retest variability, only decreases in angle are of interest here.

3.3 Level of Suppression

The proportion with each level of suppression (negligible, mild, moderate, or dense, as indicated on the completed follow-up worksheet) by treatment group at 3 months will be tabulated. The median and interquartile range for each group will be calculated and the distributions of suppression levels compared

between treatment groups at 3 months using the exact non-parametric Wilcoxon rank-sum test and two-sample Hodges-Lehmann estimator.

4 Exploratory Analyses

Exploratory analyses will describe the distribution of the outcome, or change in outcome from baseline, within each treatment group and estimate treatment group differences and 95% confidence intervals (CIs), unless specified otherwise, with the purpose of identifying hypotheses for further study. All exploratory analyses will use observed data only without imputation for missing data. Only visits occurring within the analysis window will be included. As these are exploratory analyses, no adjustment for multiplicity will be made to confidence intervals and no p-values will be reported.

Continuous outcomes will be described using mean, standard deviation, and 95% CIs. Median and interquartile range (IQR) will also be reported if the outcome distribution is asymmetric. Treatment group differences and 95% CIs will be estimated using an ANCOVA model with adjustment for baseline level of the outcome. ANCOVA is relatively robust to violations of normality and homoscedasticity[4]; however, if ANCOVA model estimation of treatment group differences is of questionable validity due to violation of normality or other assumptions, actions to mitigate the violation will be undertaken. For example, for homoscedasticity violations, an ANCOVA model allowing the treatment groups to have different variances may be used and/or large outliers may be winsorized to the 5th and 95th percentiles of the overall outcome distribution (with treatment groups combined). In cases where winsorization of outliers is used, descriptive statistics for the treatment groups will be based on the outcome distribution after winsorization, and winsorized values will be noted in table/figure footnotes. Results of mitigated and original analyses will be compared for consistency, and results of both analyses reported if there are substantial differences.

Categorical outcomes will be described by tabulating the number and percentage falling into each category by treatment group at baseline and 3 months. For binary / dichotomized outcomes, the treatment group difference in percentages with the outcome at 3 months and 95% confidence interval will be estimated by inverting Barnard's test, using the method of Agresti and Min.[2] For multichotomous outcomes, no treatment group differences, or CIs, will be estimated as the sample sizes in each treatment group are small and sample size per outcome category is likely to be sparse.

4.1 Proportion with Anomalous Retinal Correspondence

For those able to perform the test, normal retinal correspondence will be defined as correspondence status marked "normal" both when aligned and tropic, while anomalous retinal correspondence will be defined as correspondence status marked "abnormal" when aligned or tropic (or both). This outcome will be analyzed as a binary outcome.

4.2 Convergence and Divergence Fusional Amplitudes

The distribution of fusional convergence amplitude (blur point, break point, and recovery point) and fusional divergence amplitude (break point and recovery point) will be described by treatment group using medians and interquartile ranges and compared between treatment groups at 3 months using the exact Wilcoxon rank sum test. When no blur is reported for fusional convergence, the break point will be used for analysis.

4.3 IXT Symptom Survey Score

A 7-item IXT Symptom Survey will be completed by children aged 5 years and older at the enrollment and 3-month 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 at both baseline and 3 months. The symptom survey will not be Rasch-scored, and no estimates of between treatment group differences will be obtained.

Children who were less than 5 years old at the time of enrollment will not complete the survey. Instead, they will be asked a single question: "Do you have double vision (see two of things when you know there is really only one)?", with the ability to respond with "never" "sometimes", or "all the time." The distribution

of responses will be tabulated for each treatment group at both baseline and 3 months. No estimates of between treatment group differences will be obtained.

4.4 PedEyeQ Score

The PedEyeQ consists of 3 components: child, proxy, and parent. After applying existing Rasch scoring using lookup tables, the average score for each of the domains within the three components will be obtained for each of the treatment groups at 3 months using an ANCOVA model, adjusting for the baseline domain score. The treatment group difference and 95% confidence interval will be obtained from the ANCOVA model.

4.5 Binocular Diplopia by Parental Report

The number and proportion of participants with binocular diplopia at 3 months with a frequency of "more than 2 times per day" over the last week at distance and near separately (by parental report) will be tabulated by treatment group and analyzed as a dichotomized outcome.

5 Patch Wear Time

Site staff will activate and distribute occlusion dose monitors (ODMs) for participants randomized to patching (patching adherence is not applicable to the observation group). While activated, the ODM takes a temperature reading every 20 minutes, which is used to determine whether the patch was being worn at that time.

The sensors hold 133 days of data when the time interval is set at 20 minutes. If the sensor has not been read and data downloaded for a longer period than the max storage allows (133 days), the oldest data is overwritten, so only the latest 133 days will be uploaded.

For days with sensor readings, the temperature threshold for considering the patch worn will be ≥ 82°F (≥27.78 °C), versus < 82°F (27.78 °C) for not worn.[5]
For each participant in the full-time patching group, the following statistics will be reported:

- Number of participants (% of total in patching group) with ODM data.
- Distribution (median, IQR, minimum, maximum, histogram) of the percentage of days for which ODM data was obtained, including all participants randomized to full-time patching.
    - Percentage of days for which ODM data was obtained = (number of days for which ODM data was obtained (regardless of completeness))/(expected number of days of patching) * 100%.
    - Expected number of days of patching = (Date of randomization – date of 3-month visit) – 2, since expected to patch day after randomization and stop patching the day before their 3-month visit.
        - Number of participants (% of total in patching group) with no ODM data.
        - Of those with data, the distribution of the number of days patching will be split into quartiles
            - Within each quartile, the number of participants (% of total in patching group) will be calculated.
- Distribution (median, IQR, minimum, maximum, histogram) of total number of hours patched in patching group participants with ODM data
    - Total number of hours patched = (Number of readings exceeding 82 degrees Fahrenheit) x 20 minutes x (1 hour / 60 minutes)

Scatterplots, with Spearman correlation coefficients and 95% confidence intervals, will be generated to illustrate the relationship between the total number of hours patched and each outcome listed below for those with ODM data:

1. Distance control score
2. Near control score
3. Distance angle of deviation
4. Near angle of deviation
5. Suppression Score (0=negligible, 1=mild, 2=moderate, 3=dense)
6. Convergence fusional amplitudes
7. Divergence fusional amplitudes
8. Symptom Survey Score
9. PedEyeQ score (each domain)

The relationship between the total number of hours patched and the suppression score will be illustrated using a boxplot, rather than a scatterplot.

Patching adherence is categorized by the study coordinator as excellent, good, fair, or poor based on inspection of the patching calendar and discussion with the parent. This data will be tabulated for the 3-month visit, as will the reasons for any major deviations. Descriptive statistics and a boxplot for total number of hours patched as recorded with the ODM by level of patching adherence based on parent report will be obtained.

6 Association of Retinal Correspondence Testing to Other Outcomes

The relationship between retinal correspondence response and each of the following outcomes within each treatment group will be evaluated at 3 months:

1. Distance control score
2. Near control score
3. Distance angle of deviation
4. Near angle of deviation
5. Suppression Score (0=negligible, 1=mild, 2=moderate, 3=dense)
6. Convergence fusional amplitudes
7. Divergence fusional amplitudes
8. Diplopia by parental assessment (dichotomized as in section 4.5)
9. PedEyeQ score

Normal retinal correspondence will be defined as correspondence status marked “normal” both when aligned and tropic, while anomalous retinal correspondence will be defined as correspondence status marked “abnormal” when aligned or tropic (or both).

For continuous and ordered categorical outcome variables (control scores, angles of deviation, fusional amplitudes, suppression score, and PedEyeQ score), a difference in means and 95% confidence interval between participants with normal versus anomalous retinal correspondence will be estimated and illustrated using scatterplots (control scores, angles of deviation, fusional amplitudes, PedEyeQ score) or boxplot (suppression score). For categorical outcome variables (diplopia), the difference in proportions and 95% confidence interval between participants with normal versus anomalous retinal correspondence will be estimated by inverting Barnard’s test using the method of Agresti and Min.[2] P-values will not be reported.

7 Subgroup Analyses of Distance Control

The primary analysis will be repeated, stratifying by gender and race/ethnicity subgroups in accordance with NIH guidelines, to determine whether treatment effects within subgroups are consistent with overall treatment effects. These analyses will include only observed data (no imputed data). No effect modification is expected for these factors. A forest plot will be used to display the estimates of difference between treatment groups and 95% confidence intervals for each subgroup. P-values will not be reported. Subgroup analyses will only be performed for groups with a minimum of 10 observations per group.

8 Safety Analyses

All safety analyses described below will use observed data only without any imputations for missing data. Only the visits that occurred within the analysis window will be included.

8.1 Deterioration

The number and proportion of participants with deterioration, as defined below, will be tabulated for each treatment group at 3 months as a safety outcome. Treatment group differences will not be reported.

Deterioration is defined as meeting one of the following criteria at the 3-month follow-up visit:

1) Constant exotropia at least 10Δ at distance AND near by SPCT:
    a) Constant is defined as an exotropia present throughout the examination and confirmed by cover/uncover testing performed at least 3 different times over the course of the exam (including during assessment of IXT control).
2) Decrease in near stereoacuity of at least 2 levels from baseline, or to nil, measured using the Randot Preschool Stereotest.

8.2 Reduction of Distance Visual Acuity

Any cases of 1) interocular difference ≥0.3 logMAR and 2) reduced visual acuity in best refractive correction (≥0.3 logMAR) in either eye will be tabulated by treatment group at 3 months.

8.3 Diplopia

At 3 months, frequency of diplopia by child survey will be tabulated for each treatment group as “Never”, ‘Sometimes”, or “All the Time”. By parent survey, the responses will be tabulated by treatment group as “No”, “2 times or less per day”, and “More than 2 times per day”.

9 Additional Tabulations

The following additional tabulations will be performed:

- A flow chart accounting for all participants according to treatment group
- Number and percentage completing the visit by treatment group
- Baseline demographics and clinical characteristics overall and by treatment group at randomization
- Protocol deviations by treatment group

References

1. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006;93(3):491-507. doi:10.1093/biomet/93.3.491
2. Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*. Sep 2001;57(3):963-71. doi:10.1111/j.0006-341x.2001.00963.x
3. Hatt SR, Leske DA, Liebermann L, Mohney BG, Holmes JM. Variability of angle of deviation measurements in children with intermittent exotropia. *J AAPOS*. Apr 2012;16(2):120-4. doi:S1091-8531(12)00074-2 [pii]
10.1016/j.jaapos.2011.11.008
4. SF O, J A. Parametric ANCOVA and the Rank Transform ANCOVA When the Data are Conditionally Non-normal and Heteroscedastic. *J Educ Stat*. 1984;9(2):129-149.
5. Wang J, Jin J, Malik A, et al. Feasibility of monitoring compliance with intermittent occlusion therapy glasses for amblyopia treatment. *J AAPOS*. 2019;In Pressdoi:10.1016/j.jaapos.2019.04.009