Supplementary Online Content 4

The Multicenter Uveitis Steroid Treatment (MUST) Trial and Follow-up Study Research Group. Association between long-lasting intravitreous fluocinolone acetonide implant vs systemic anti-inflammatory therapy and visual acuity at 7 years among patients with intermediate, posterior or panuveitis. *JAMA*. doi:10.1001/jama.2017.5103

Statistical Analysis Plan for MUST Follow-up Study

1

Association Between Long-lasting Fluocinolone Acetonide Intravitreous Implant vs Systemic Anti-inflammatory Therapy and Visual Outcomes at Seven Years Among Patients With Intermediate, Posterior, or Panuveitis

Statistical Analysis Plan

Version 3
19 March 2017

2

3

The primary outcome measurement in the MUST Trial Follow-up Study is the change in BCVA from enrollment to 7 years of follow-up. The statistical methods used to analyze the long-term follow-up in the continuation study will be the similar to those outlined in the original MUST Trial. For the MUST Follow-up Study, an observational study, the primary analysis still will be according to the original randomization group, following the principles of Intent to Treat (ITT) in that all patients will be grouped according to their assigned therapy regardless of the actual treatment received. This allows us to answer the real-world question of the long-term prognosis of an individual undergoing an intended course of therapy. Secondary analyses will be based on the treatment received, and the ITT analysis will be used as a point of departure from which a causal analysis based in time-varying covariates and propensity scores to account for selection and dropout effects will be built. The relation between the ITT analysis and the causal modeling will be of interest in its own right, but our principal focus will be on understanding the long-term trajectories via the analysis of the observational data. One of the treatment-based analyses will focus upon those individuals who receive treatment according to the protocol specifications. Individuals (or eyes) will be included in this analysis if they received their assigned treatment within six months of randomization. For individuals (or eyes) assigned to systemic treatment, follow-up will be censored at time of implantation, if an implant is placed. All primary analyses will be replicated by at least two different analysts.

As during the trial phase of MUST, evaluation of continuous outcomes over time (such as change in BCVA) and binary outcomes over time (such as normal/abnormal IOP) will use a repeated measures analysis with Gaussian or logit links and accounting for the nested correlations between observations over time and (when applicable) between eyes of the same participant.[1] Both absolute (percent) and relative (odds ratios) comparisons will be presented for binary outcomes using unadjusted and adjusted logistic models, respectively. The primary analysis will be based upon generalized estimating equations (GEE) since our goal is to summarize the marginal (i.e. population average) effects. Sensitivity analyses using mixed effects models will also be explored; however, since the goal is the population as opposed to individual level effects, they will primarily be used to assess the impact of missing data, incorporation of clinic (as a random effect), and alternate covariance structures. The mixed effects analysis also will allow for the exploration of the effect of clinic on the results. For short to moderate term follow-up, a saturated mean model, including visit and visit by treatment interaction terms, will be used. An unstructured covariance matrix will be used to model the within-eye repeated measurements augmented by random effects to induce cross-sectional between-eye associations in the clinical trials. However, the unstructured covariance model will not be feasible for the long-term follow-up study, because a large number of parameters need to be estimated. Therefore, we will replace it by a Toeplitz covariance structure or, for the mixed effects model, a structure composed of a first-order, auto-regressive process (an AR(1) model) along with a random intercept. These structures allow for correlation to decrease with increasing time-separation.

Evaluation of risk factors for time-to-event outcomes such as incidence of ME, cataracts, IOP elevation or glaucoma, as well as time from ME to remission will be performed using Cox

proportional hazards regression as well as parametric time-to-failure models, such as gamma models[2], including a random effects term to account for the between eye correlation. Implementation of these models allows for clustering (within participants and within eyes), assessment of recurrent events, and incorporation of time-dependent covariates. When the relevant follow-up time in the analysis reflects the clinical time scale (e.g. time since diagnosis of uveitis), it is necessary to incorporate the prevalent cases (longstanding diagnosis of uveitis) into the analysis. To accommodate the prevalent cases we will use the staggered entry technique[3] which is one method of adjusting for potential survival bias.
The analysis compares the event rates among participants with similar duration of disease and then combines over these comparisons. Event rates for multiple recurring events, e.g. the number of adverse events, will be modeled using Poisson regression or Negative Binomial regression , including a random effects term to account for the between eye correlation.[2]

Cross-sectional comparisons between subgroups (e.g. those who did and did not complete follow-up) will be performed. For person-level characteristics, χ^2 tests or Fisher's exact test and Wilcoxon rank-sum tests will be used to compare categorical and continuous variables, respectively. For eye-level characteristics, mixed effects models will be used.

All analyses will be performed both unadjusted and adjusted for potential confounders. Effect modification due to factors such as disease location, systemic disease, gender and race also will be explored when appropriate. Robust standard errors will be computed using statistical program-based approaches when available and a bootstrap with the individual as the sampling unit, when a pre-programmed approach is not available. The bootstrap was primarily used to adjust for between-eye correlation for ocular outcomes.

The bootstrap routine for estimating standard errors, confidence intervals, and p-values is as follows. First, the original estimates will be obtained by analyzing the trial and follow-up study data adjusting for longitudinal within-eye measurements assuming a working independence structure for the between-eye correlations, i.e. treating each eye as an independent observation. A total of 5000 iterations will be used to calculate the bootstrap estimates. The sampling unit is the individual (as opposed to the eye) to preserve the between-eye correlation structure. Within each subgroup defined by treatment and disease laterality (unilateral systemic, bilateral systemic, unilateral implant, and bilateral implant), we will sample with replacement to create a new cohort of the same size. The resulting datasets will maintain sample size, both in terms of individuals and eyes, found within each treatment group in the original dataset. We will then apply the model developed for the original dataset to each of the bootstrap samples. The standard deviation of the bootstrap estimates will be used as the standard error for the original estimates. Assuming a normal distribution, t-test statistics, p-values, and 95% confidence intervals may be constructed around the original parameter estimates using the bootstrap standard error. Alternately, the 95% confidence interval may be constructed by extracting the 2.5- and 97.5-percentiles from the bootstrap estimates.

The follow-up study focuses on a set of primary research questions and related analyses. All tests will be considered statistically significant at the 0.05 level. However, a large number of comparisons are planned for secondary outcomes and caution is needed in the reporting of interpretation of these results. As recommended by Wang et al.[4] our primary focus for these outcomes will be on the parameter estimates and confidence intervals rather than p-values. Several methods of adjusting p-values for multiple comparisons exist, however no clear consensus as to the most appropriate method is available and it is difficult if not impossible to quantify the number of comparisons. In general, issuing cautions is sufficient, but for identifiable and related sets of estimates we will do adjustments. We expect that related sets of estimates will have a high positive correlation, making a Bonferroni correction extremely conservative. Therefore, we will also explore using an estimate of the covariance matrix for these related sets based upon a bootstrap approach and also estimate the null distribution of the minimum p-values for the multivariate distribution of Z-scores using a global null hypothesis permutation distribution or the multivariate normal cumulative distribution program is R. A variety of sensitivity analyses will be performed in order to determine the potential for bias due to modeling assumptions, missing data, and potential biases (especially for epidemiologic analyses). Transformations (e.g., log) of continuous outcomes will be considered when violations of the Gaussian assumption occur.

Single imputation techniques including 'best' and 'worst' case scenarios will be used to identify the magnitude of the potential impact of missing data. For example, we will calculate the magnitude of the effect sizes that would need to be observed among those individuals with missing data in order to change our inference or conclusions. Likelihood based methods such as mixed effects models that are robust to data that is missing at random will be explored.[5] In addition, multiple imputation (MI) and pattern mixture approaches will be used to assess the impact of missing data.[5-6] Specifically, we will retain all features of the primary analysis other than how missing values for visits beyond the last one with a measured value are handled. As is implicit in the primary analysis where missing data indicators are used (e.g. "." in SAS, "NA" in R), we will treat all missing values, before the last measured one, as Missing at Random (MAR). We will impute other missing values to generate 10 pseudo-complete records for each such individual by sampling from a joint predictive distribution for the missing data given the observed data. We will use a covariance matrix equal to the estimated covariance matrix from the primary analysis, but will "take control" of the prediction mean. Varying the mean of the predictive distribution allows us to assess the sensitivity of our results to a variety of missing data scenarios. The pattern mixture model approach stratifies participants on their pattern of missing data, estimates stratum-specific parameters and then combines estimates over strata using inverse variance weights. The approach is similar to stratified analysis to adjust for potential confounders and allows for comparison of stratum-specific estimates. We stratify by three patterns: complete data, at least one internal measurement missing, and closeout weight missing (along with any other missingness).

The presence of time-varying confounders that are themselves affected by prior levels of exposure may produce large biases in the estimation of causal effects using standard statistical analyses. A classic example, demonstrated by Cole et al. [7] shows that adjusting for time-varying CD4+ T cell count can dramatically reduce the estimated benefit of HAART on time to AIDS or death. One potential example in the MUST Trial Follow-up Study would be adjusting for cataract status when estimating the time to vision loss for participants receiving regional corticosteroid treatments, since these treatments are known to cause cataracts. Another example influencing participant based outcomes is utilization of immunological agents with systemic corticosteroid treatments. Marginal structural models can be used to correct both for bias induced by a variable affected by exposure

and for bias induced by loss to follow-up.[8] The technique employs time-varying inverse probability weights in the place of standard covariate adjustments. We will explore the presence of time-varying confounding by comparing standard Cox proportional hazards models with marginal structural models.

Document revision history

Version 1 of this statistical analyses plan was developed from the MUST Trial (Version 2.4) and Follow-up Study (Version 4.4) Protocols and was submitted as part of the original JAMA submission. Substantive clarifications to the original statistical analysis plan were made in response to reviewer queries in Version 2 (1 March 2017). All revisions are captured in the tracked change versions of the documents archived at the Coordinating Center.

1. Analysis datasets (p.2)
Clarification – The primary analysis will be based upon the principles of intention to treat (ITT)

2. Analytic models (p.2)
Clarification – Generalized estimating equations will be used for the primary analyses with secondary analyses based upon mixed effects models.

Clarification – Specified that both absolute and relative effects would be calculated for binary outcomes in response to the JAMA review. Additional details of the absolute analysis are provided on p. 3.

3. Bootstrap analysis (p. 3)
Clarification – Additional details on the bootstrap methods to adjust for between-eye correlation and to estimate absolute effects for binary outcomes were added at the request of the JAMA review.

4. Missing data (p. 3)
Clarification – Additional details of the missing data analyses including ‘best’ and ‘worst’ case scenarios as well as mixed effects models were added.

5. References (p.6)
Clarification – References were corrected and a reference list added to the file.

Version 3 (18 March 2017). incorporates additional clarifications in response to the second round of review All revisions are captured in the tracked change versions of the documents archived at the Coordinating Center.

1. **Treatment-based analyses (p.2)**
   Clarification – Specifics of the criteria for the per-protocol treatment-based analysis were added.

2. **Role of mixed effects model (p.2)**
   Clarification – Exploration of the effect of clinic will be performed using mixed effects models

3. **Cross-sectional comparisons of characteristics (p.3)**
   Addition – Specifics for cross-sectional comparisons are specified.

4. **Level of statistical significant (p.3)**
   Clarification – The statistical level of significance (0.05) was added to the discussion of multiple comparisons.

192
193 References
194
195
196 1. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of longitudinal data. New York: Oxford
197 University Press.

198 2. Keene ON, Jones MR, Lane PW, Anderson J. Analysis of exacerbation rates in asthma and
199 chronic obstructive pulmonary disease: example from the TRISTAN study. Pharm Stat
200 2007; 6(2): 89-97.

201 3. Tarwarter PM, Mellors J, Gore ME, Margolick JB, Phair J, Detels R, Munoz A. Methods to
202 assess population effectiveness of therapies in human immunodeficiency virus incident
203 and prevalent cohorts. Am J Epidemiol 2001; 154(7): 675-81.

204 4. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Reporting of subgroup analyses in
205 clinical trials. N Eng J Med 2007; 357: 2187-2194.

206 5. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: John Wiley &
207 Sons. 2002

208 6. Daniels MJ, Hogan JW. Missing data in longitudinal studies: strategies for Bayesian modeling
209 and sensitivity analysis. Boca Raton: Chapman & Hall/CRC 2008.

210 7. Cole SR, Hernan MA, Margolick JB, Cohen MH, Robins JM. Marginal structural models for
211 estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count.
212 Am J Epi 2005; 162: 471-8.

213 8. Hernan MA, Brumback B, Robins JM, Marginal structural models to estimate the joint causal
214 effect of non-randomized treatments. J Am Stat Assoc 2001; 96 (454): 440-48.

215