WASH Benefits Kenya Primary Outcome Analysis Plan Update

October 12, 2016

Background

The WASH Benefits study published a detailed design and analysis protocol for primary outcomes in 2013 (Arnold *et al*. 2013). The protocol included nearly all details of the analysis plan for primary outcomes: diarrhea and length for age Z-scores (LAZ). Here, we pre-specify updates to the analysis plan for the primary outcomes in the Kenya trial. We have also included the secondary outcomes stunting and severe stunting in this analysis plan (stunting defined as LAZ < -2, severe stunting defined as LAZ < -3) because we plan to report them in the primary outcome paper for Kenya. Below, we have listed all changes from the 2013 analysis plan along with a rationale. Please refer to the design and rationale protocol for specific details about outcome definitions, measurement methods, and hypothesis tests (Arnold *et al*. 2013).

Primary Analysis: changes and additions from the original protocol, along with rationale

- In the original protocol, we specified that we would calculate 95% confidence intervals using a nonparametric bootstrap for both the unadjusted and adjusted estimates of our parameters of interest. At the time we expected to need to use a nonparametric bootstrap to calculate confidence intervals in the adjusted analysis and wanted to keep a parallel structure across the unadjusted and adjusted analyses. Advances in the tmle R software package since the original protocol was published now readily implement influence curve-based standard errors and confidence intervals that account for clustered outcome data (Gruber and van der Laan 2012). We will use influence curve-based standard errors to calculate 95% confidence intervals in the adjusted analysis (with pairs defining independent units). Therefore, we elected to use the paired t-test for unadjusted LAZ comparisons and the Mantel-Haenszel prevalence ratio and difference for unadjusted diarrhea and stunting comparisons (Donner and Klar 1993), with randomization block (groups of usually 9 clusters in the same administrative location) defining matched pairs/stratification. The confidence intervals based on the paired t-test, Mantel-Haenszel estimator, and influence curve are asymptotically equivalent to the nonparametric bootstrap but are computationally much simpler.
- We will base conclusions on the unadjusted models unless the adjusted models have better precision (smaller standard errors).
    - In the original protocol, we specified that we would calculate one-sided *P*-values because we expected the interventions to only be beneficial. After a September 2014 collaborative meeting with investigators from the related SHINE trial (NCT01824940) and the published results from a study in Kenya that found iron fortification could increase

enteric pathogen abundance in young children (Jaeggi 2015), we decided to take a conservative approach and report two-sided *P*-values for all hypothesis tests.

- As specified in the original protocol, we will conduct the adjusted analyses using targeted maximum likelihood estimation (TMLE) on child-level outcomes. TMLE is implemented in the `tmle` package in the R statistical software (Gruber and van der Laan 2012). TMLE enables the use of the SuperLearner[5] ensemble machine learning algorithm (van der Laan, Polley, and Hubbard 2007) to flexibly adjust for the pre-specified covariates in our original protocol. We will include the following model library in the ensemble [`SuperLearner` R package arguments in brackets]: the simple mean [`SL.mean`], main effects generalized linear models [`SL.glm`], Bayesian generalized linear models [`SL.bayesglm`] (Gelman *et al.* 2008), generalized additive models degree 2 [`SL.gam`] (Hastie and Tibshirani 1990), and penalized elastic net regression [`SL.glmnet`] (Friedman, Hastie, and Tibshirani 2010). We chose this model library because it enables flexible relationships between covariates and the outcome but does not include algorithms with internal bootstrapping or bagging (e.g., RandomForest (Breiman 2001)), which can lead to overfitting the data.
- The original protocol stated that within study compounds, health outcome data (including diarrhea) would be collected from all children <36 months at enrollment. Due to time constraints during data collection at enrollment and logistical considerations for the follow-up visits, data on health outcomes were collected from a subset of children <36 months in study compounds as follows:
    - Enrollment: all children 18-27 months in the compound (these children were enrolled in the parasite substudy); if there were no children eligible for the parasite substudy then health outcome data were collected for one child <36 months living in the study household (index child if born); if there were no children <36 months in the study household and no parasite children in the compound, then health outcome data were collected one child <36 months living in another household in the compound
    - Year 1 and Year 2 follow-up visits: all children <36 months in the study household
- Because this data collection process generates age distributions that differ across survey rounds, and because there are relatively few non-index children for the diarrhea analysis (since we pre-specified that we would not include new births after the index children as they might not have had sufficient exposure to the interventions by the follow-up visits and index children are not very likely to have older siblings who are still <36 months), we will use the following samples of children for the diarrhea analyses:
    - Enrollment: all non-index children <36 months with health outcomes data (to avoid the spike in the age distribution at 0 months due to preferentially sampling index children who were born between the census and the baseline data collection)
    - Year 1 and Year 2 follow-up visits: index children only (including twins), from the day anthropometry was collected

- Due to resource constraints, we did not administer the uptake module in the passive control arm during the follow-up surveys so we cannot compare behaviors between the active and passive control arms. We did collect data on primary and some secondary outcomes in the passive control arm. Specifically, the passive control arm was not included in the environmental enteropathy or parasites subsamples, but we have data on growth, health outcomes, and child development for the passive control arm.
- We have decided to drop several covariates originally intended to be included in the adjusted analyses for the following reasons: We removed administrative location because it does not add information once we have already included fixed effects for the geographic stratification. We removed time between intervention delivery and measurement because this characteristic could not be measured in control clusters (no intervention) and because it was post-randomization. We removed maternal literacy from our set of pre-specified adjustment covariates because it is highly correlated with maternal education. We expanded our age range for number of children in the household to include children 15-17. All maternal, household, and compound characteristics were measured at enrollment (before randomization) except for maternal height, which was measured at the 1 or 2 year follow-up visits to avoid the influence of pregnancy, since mothers were pregnant at enrollment. The final adjustment covariate list is below:
    - Month of measurement, to account for seasonal variation
    - Household hunger scale (3 categories: little or no insecurity, moderate, and severe insecurity) (Ballard et al. 2011)
    - Child age (days)
    - Child sex
    - Mother's age (years)
    - Mother's height (cm)
    - Mother's education level (incomplete primary, complete primary, any secondary)
    - Number of children < 18 years in the household
    - Number of individuals living in the compound
    - Distance (in minutes) to the household's primary water source
    - Field staff identification number
    - Housing materials (floor, roof) and household assets
        - Assets measured. Has: electricity, radio, television, mobile phone, clock, bicycle, motorcycle, stove, gas cooker, car. Number of: cows, goats, dogs, chickens.
- We will pre-screen covariates to assess whether they are associated with each outcome prior to including them in adjusted statistical models (Pocock *et al*. 2002). We will use the likelihood ratio test to assess the association between each outcome and each covariate and will include covariates with a p-value < 0.2 in the adjusted analysis. We will also exclude any covariates that have very little variation in the study population (prevalence <5%).

Tests for potential between-cluster spillover effects (contamination), changes and additions from the original protocol, along with rationale

In Appendix 5 of the study's protocol, we outlined a nonparametric test for between-cluster spillover effects (i.e., contamination) (Arnold *et al*. 2013). Although we expect such contamination to be very unlikely due to the low population density and the proportion of compounds enrolled in the trial, following our original protocol we will conduct formal tests to rule out such effects, which could bias the measures of intervention toward the null (by making outcomes across arms appear more similar than if arms were independent). Below, we have added details to the analysis plan for this statistical test that were not included in the original protocol.

- We will assess potential between-cluster spillovers for the primary outcomes (child diarrhea using all data on index children at the year 1 and 2 follow-up visits and length-for-age at year 2).
- The original statistical test incorporated information about population density. Although we collected GPS coordinates and data on the number of residents for each of the 174,000 compounds visited during the census to determine eligibility, we do not have the resources it would take to convert these data into population density measures in advance of the primary analysis, so the models for between-cluster spillovers will not directly adjust for local population density.
- Consistent with our original analysis plan, which defined between-cluster spillovers as a function of the number of treated compounds within specific distances, we will estimate the association between each primary outcome (Y) among active and passive control households (T=C) and the proportion of study households within 2 km of each control household that are exposed to each of the interventions (N^T). The threshold of 2 km was chosen based on the distribution of distance to the nearest treatment household from control households: roughly 75 percent of control households are within 2 km of their nearest treatment household. To test for the presence of between-cluster spillovers, we will use a clustered permutation test for sets of interventions to test the following null hypothesis:

$$\Psi = \mathrm{E}[Y \mid T=\mathrm{C}, X, N^T \leq n^T_{20}] - \mathrm{E}[Y \mid T=\mathrm{C}, X, N^T \geq n^T_{80}] = 0,$$

where

- N^T is the proportion of study households within a 2 km radius of each active control household which are assigned to the following sets of intervention arms {Water, Water+Sanitation+Handwashing (WSH), Nutrition + WSH}, {Sanitation, WSH, Nutrition + WSH}, {Handwashing, WSH, Nutrition + WSH}, {Nutrition, Nutrition + WSH}
- n^T_{20} is the 20th percentile of the distribution of N^T
- n^T_{80} is the 80th percentile of the distribution of N^T

We will conduct both an unadjusted test and a test adjusting for baseline covariates (X). For the adjusted test, we will fit $E[Y \mid T, X, N^T]$ using a repeatable data-adaptive algorithm (van der Laan, Polley, and Hubbard 2007).

If between-cluster spillovers are present, we would expect Ψ to be greater than 0 for length-for-age Z-scores and less than 0 for diarrhea prevalence (i.e., households that are near more treatment households should benefit from intervention more than households that are near fewer treatment households). We are pre-specifying the 20th and 80th percentiles of the distribution of N^T because they are close to the lower and upper range of the distribution (i.e., close to the extremes), but still have sufficient observations to ensure that we can estimate the quantities with reasonable precision.

- If we find evidence of between-cluster spillover effects, we will interpret our intention-to-treat estimates as lower bounds of the treatment effects under the assumption that spillovers are positive (i.e., in the same direction as treatment effects). In this case we may also test for spillovers in the following behavioral indicators to help explain the mechanism of spillover:
  - Households have chlorinated drinking water (measured by residual free chlorine)
  - The participant reports that the youngest child's most recent defecation was either directly into the latrine or the feces were captured by the potty and disposed of into the latrine or buried
  - Households have access to an improved latrine
  - Households have at least one handwashing station with soap and water present
  - The participant reports hearing any messages on infant/child nutrition and or Mwanzabora (lipid based nutrient supplement)

References

Arnold BF, Null C, Luby SP, *et al*. 2031. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale. *BMJ Open* 3(8):e003476.

Ballard T, Coates J, Swindale A, Deitchler M. 2011. Household Hunger Scale: indicator definition and measurement guide [Internet]. FANTA. Available from: http://www.fantaproject.org/sites/default/files/resources/HHS-Indicator-Guide-Aug2011.pdf

Breiman L. Random Forests. Mach Learn 2001; 45(1):5–32.

Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomization trials. J Clin Epidemiol 1993; 46(2):123–31.

Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2008; 2(4):1360–83.

Gruber S, van der Laan M. tmle: An R Package for Targeted Maximum Likelihood Estimation. J Stat

Softw 2012; 51(13):1–35.

Jaeggi T, Kortman GAM, Moretti D, *et al*. 2015. Iron fortification adversely affects the gut microbiome, increases pathogen abundance and induces intestinal inflammation in Kenyan infants. *Gut* 64(5):731–42.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 21(19):2917–30.

van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1): 1–21.

Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. Springer; 2002.