# Sanitation, Water, and Instruction in Face-washing for Trachoma

# Statistical Analysis Plan

Confidential

Version 1.9, 29 October 2020

F.I. Proctor Foundation for Research in Ophthalmology

# 1 Introduction

This document (Statistical Analysis Plan, SAP) describes the planned analysis and reporting for the **Sanitation, Water, and Instruction in Face-Washing Trial,** University of California, San Francisco, J. Keenan, PI. It includes specifications for the statistical analyses and tables to be prepared for the final Clinical Study Reports.

The WUHA trial is a community-randomized trachoma control trial comparing previously treated communities which receive water, hygiene, and sanitation interventions with previously treated communities which did not receive these interventions. The TAITU trial is a community-randomized trachoma control trial comparing (A) repeated antibiotic treatments targeted to chlamydia-positive 0-5 year old children, (B) mass azithromycin distributions given to the entire community, and (C) delayed antibiotic treatment. All SWIFT clusters received 8+ years of treatment prior to the study period.

The content of this Statistical Analysis Plan meets the requirements stated by the US Food and Drug Administration (Department of Health and Human Services, Food and Drug Administration, 1998) and conforms to the American Statistical Association's Ethical Guidelines (American Statistical Association, 1999).

The following documents were reviewed in preparation of this Statistical Analysis Plan:

- Trial Manual of Operations, and Proposal
- ICH Guidance on Statistical Principles for Clinical Trials (Department of Health and Human Services, Food and Drug Administration, 1998)
- Statistical Analysis Plan (prepared by T. Porco), Trachoma Amelioration in Amhara II (T. Lietman, PI)

The planned analyses described in this Statistical Analysis Plan will be included in future manuscripts. Note, however, that exploratory analyses not necessarily identified in this Statistical Analysis Plan may be performed to support the analysis. All post-hoc or unplanned analyses which have not been delineated in this Statistical Analysis Plan will be clearly documented as such in the final Clinical Study Report, manuscripts, or any other document or submission.

Finalization of this document will take place prior to the final study visit the final version of this document will be numbered 2.0. All subsequent changes will be indicated by detailed change log in an Appendix.

J. Keenan, D. Fry, T. Lietman, and T. Porco have contributed to this plan.

# 2 Design

The study consists of 2 complementary cluster randomized trials, designated as the "WUHA" or "WASH trial" and "TAITU" or "targeted antibiotics trial." The design of the trials is described fully in the Proposal and the Manual of Operations. Masking, selection of communities, inclusion criteria, exclusion criteria, and informed consent are discussed in those documents.

## 2.1 Outcomes

The design is community-randomized, based on school districts. Details are provided in the Protocol (WUHA Research Plan).

Interventions and monitoring will take place only in a subset of individuals within each school district, as indicated in the Protocol. The primary outcome variable is the prevalence of *Chlamydia trachomatis* in children, as assessed from PCR of conjunctival swabs. The age ranges, sampling frame, and grading or laboratory methods are described in the Protocol.

## 2.2 Sample size

The two trials have different sample sizes, based on the expectation that the effect size for the WASH intervention will be lower over the short time period of the trial than will be the effect size of the targeted antibiotic intervention.

The sample size for the WASH trial is based on the primary outcome of Specific Aim 1: the cluster-level prevalence of ocular chlamydia. Power is reported for secondary aims and for the other Specific Aims (Appendix).

The sample size has been calculated using the formula for a two-sample T-test (as given in Section 3.2.1 of Chow et al (2007). While the power has been computed using the T-test formula, the statistical analysis will not be based on a simple T-test, but on a regression model which should yield greater power. We assume a two-sided alpha of 0.05, a power of 0.8, an effect size of 0.08 (eight percent), and a standard deviation of 0.10 to obtain a sample size of 22 communities per arm, or 44 communities in all. The standard deviation of 0.10 was obtained by (1) calculating the standard deviation from previously treated communities in the TANA study (Stoller et al, 2011), (2) inflating the standard deviation slightly to correct for the fact that the TANA study took place in somewhat larger communities than the proposed WUHA trial (details are in the Appendix to this SAP), and finally (3) reducing the standard deviation, because we assume the baseline prevalence has a correlation coefficient of 1/3 with the final value of the prevalence and thus explains approximately 9% of the variance in the final outcome. No assumptions have been made that any communities will be lost to followup. The sample size plan is based on a cluster level analysis and therefore reflects the clustered nature of the design.

Due to a drought in the study area, no more than 40 school districts with communities who had an eligible water point could be identified. Including 20 clusters per group provides 79% power

(i.e., 3% less power than the originally planned sample size) to detect an 8% effect size, given the same assumptions described above.

As a guide to the sample size for Specific Aim 2, we proceed as follows. Using similar methods and rationale, 16 study clusters per treatment arm would provide at least 80% power to test a noninferiority margin of a 10% difference between the study arms, assuming a standard deviation of 0.10, a correlation of 0.50 between the baseline and final prevalence, and an alpha of 0.05. The higher correlation between baseline and final prevalence is based on the Stoller 2011 study; this is reasonable since study clusters would be expected to be more correlated over this 24-month study than in the 36-month WASH trial described above. For Aim 2B, we assume 30 children per community, a loss to follow-up of 15%, a community-level ICC of 0.01 for ocular chlamydia infection based on our own previous unpublished studies in Ethiopia, a final chlamydia prevalence of 15% in the control arm, and a correlation of 0.5 between individual-level baseline and final chlamydia status. Under these assumptions, 16 communities per arm would provide greater than 80% power to detect an absolute effect size of 7.5% between the two arms.

## 2.3 Randomization

Procedures for site selection, community eligibility criteria, and informed consent procedures are discussed in the Protocol and the Manual of Operations. Application of these procedures must yield a set of no fewer than 40 suitable communities. The sentinel study community will be chosen prior to randomization, using the procedure described in the Manual of Operations.

**WUHA:** Given a set of 40 school districts with communities which are suitable for inclusion, a simple random sample will be drawn of 20 communities to be assigned to the WASH intervention arm and 20 to be assigned to the WASH control arm

**TAITU:** Given a set of 48 communities suitable for inclusion, a simple random sample will be drawn of 16 villages to be assigned to the targeted antibiotic arm, 16 villages to be assigned to the MDA arm, and 16 villages to be assigned to the delayed treatment arm. Details of randomization are provided in the Appendix to this Statistical Analysis Plan.

# 3 Primary hypothesis tests

Procedures for handling missing data, checking model assumptions, and conducting secondary analyses are provided in the Appendix.

## 3.1 WUHA (Water, Sanitation and Hygiene) Specific Aim

**Primary outcome** . The following information is anticipated:

1. $Y_{it}$, the prevalence of infection in village *i* at time *t*, where *t*=1,2, or 3 (year 1, 2, or 3). This is the outcome variable.
2. $P_i$, the baseline prevalence in village *i* (which could have been denoted $Y_{i0}$). As discussed elsewhere, this is the prevalence in children aged 0-5 as determined by PCR testing. The age limit of 5 is inclusive.
3. $X_i$, the treatment covariate for each village, 0 for control group, and 1 for intervention group.
4. $T_{it}$, the time the infection data was collected in village *i* at time *t*, in continuous months after baseline.

We wish to use all three time points as part of the outcome using a pooled or longitudinal scheme. Several possibilities must be considered: (a) the effect of the intervention is fast, and the village-specific prevalences in the treatment arms are lower than baseline and lower than the control groups, and show no further trend: the slopes are the same, but the mean is lower in the treatment group than the control group, and (b) the effect of the intervention is gradual, with the difference between the control and intervention group becoming larger over time.

We propose to fit a linear mixed effects regression model to the village-specific prevalences over time. A random intercept term will be included because successive measurements on the same village may show statistical dependence; the fixed effects will be modeled according to the following equation:

$$Y_{it} = \beta_0 + \beta_1 P_i + \beta_2 T_{it} + \beta_3 X_i + \beta_4 T_{it} X_i$$

Estimation will be conducted by restricted maximum likelihood. We propose to include the interaction term only if it is statistically significant. Testing will be two sided at an alpha level of 0.05, based on the value of $\beta_3$, testing the difference between the control and intervention communities. If the interaction term is included, then the deviance statistic contrasting the model with all terms with the model without $\beta_3$ and $\beta_4$ will be used.

Significance testing will be conducted by permutation testing at the village level, and will reflect the entire fitting algorithm, preserving the overall alpha level at 0.05. We anticipate modestly higher power than used for planning, since data from all three follow-up visits will be used, and not just the data at the final visit.

Diagnostics, alternatives, and sensitivity analyses are discussed in the Appendix.

## 3.2 Additional Secondary analyses, WUHA (Water, Sanitation, and Hygiene Aim

Diagnostics, alternatives, and sensitivity analyses are discussed in the Appendix.

3.2.1. Village-level *Chlamydia* load outcome.

The analysis is proposed to be identical to that of the ocular chlamydia outcome in §3.1, except that a village-specific index of chlamydia load at baseline and at follow-up times is used instead of prevalence. The analysis is two-sided at an alpha of 0.05.

This is a prespecified secondary analysis and will be reported as such.

3.2.2. Follicular trachoma scores, longitudinal analysis.
Follicular trachoma scores (using the 5 level system) will be modeled longitudinally using linear mixed effects regression. Scores are indexed by participant, grader, visit, and village. Participant and village will be modeled as random effects; grader will be modeled as a fixed effect. We will use an AR(1) correlation structure.

3.2.3. Inflammatory trachoma scores, longitudinal analysis.
Inflammatory trachoma grades will be modeled in the same way as in Section 3.2.2.

3.2.4. Chlamydial load, individual level analysis
Quantitative PCR results at the individual level will be modeled using standard procedures for semi-continuous variables.

3.2.5 Ocular chlamydia, age 6-9 or older
Longitudinal analysis of ocular chlamydia in the age 6-9 group will be modeled at the village level. A similar analysis will consider the 10 and over segment of the population.

3.2.6 Nasopharyngeal macrolide resistance
Nasopharyngeal macrolide resistance in age 0-5 will be modeled at the village level, using treatment arm as a covariate.

3.2.7 Clean faces.
The proportion of the population with clean faces will be compared (at the village level) between the two groups, using ANCOVA with baseline values and treatment arm as covariates.

3.2.8 Anthropometry
Longitudinal analysis of anthropometric measurements between the two arms will be conducted using growth curve models.

3.2.9 Clinical trachoma
We anticipate having the following information available. For each child in the sampling frame, we will have a binary improvement score, based on baseline and follow-up photography. We propose to conduct clustered logistic regression (taking into account the clustered nature of the

design) using village assignment as the predictor. We will estimate the log odds of the treatment effect. Significance testing will be conducted at 0.05 based on Monte Carlo permutation testing.

3.2.10. Soil-transmitted helminth prevalence (village-level analysis)
The prevalence of soil transmitted helminths will be compared between the WUHA treatment arms in a linear mixed effects regression as described above for ocular chlamydia.

3.2.11. Soil-transmitted helminth density (individual level analysis)
Quantitative soil transmitted helminth results at the individual level will be modeled using standard procedures for semi-continuous variables.

3.2.12. Prevalence of chlamydia antigen positivity from serological tests (village-level analysis)
The prevalence of soil transmitted helminths will be compared between the treatment arms in a linear mixed effects regression as described above for ocular chlamydia.

3.2.13 Passive surveillance
From previous work by our group performing passive surveillance work, we have learned that individual communities access health care at health posts and health clinics in a heterogenous way. It is unclear exactly which factors are important, but we speculate that distance to the health post, presence of private health clinics, and relative poverty level may all play a role. In order to account for this heterogeneity, we will attempt to collect passive surveillance data for the calendar year before the trial starts in order to have a baseline measure of health post usage that accounts for the seasonality of different diseases.

We will count the number of children aged 0-5 years, children aged 6-9 years, and individuals ≥ 10 years per study cluster who present to a government health post in the study area during the trial, and compare the treatment arms with negative binomial regression, using the age-stratified census population as the offset and adjusting for the number of individuals seen in the previous year. We will compare the number of (1) infectious disease visits, (2) diarrhea visits, (3) respiratory infection visits, and (4) malaria visits in each age strata between treatment arms.

Univariate analyses and multivariate analyses will be conducted. Specific covariates of interest are treatment arm, individual age, the presence of clean faces, the presence of latrines, and access to water.

## 3.3 TAITU: Targeted Antibiotics Specific Aim

**TAITU-A Primary Outcome**
The primary outcome will be the cluster-specific prevalence of ocular chlamydia among individuals aged 8-12 years, compared between the targeted azithromycin arm and the mass azithromycin arm in a non-inferiority analysis.

The statistical approach will be to conduct linear regression of the final 24-month values for each village, using the baseline values and treatment assignment as covariates.

Evidence of outliers or strong departure from normality will necessitate the use of robust regression. Permutation based P-values will be reported; all regressions conducted will always be reported.

A noninferiority test will be conducted, using a noninferiority margin of 0.1, based on a two-sided 95% confidence interval for the regression coefficient for assignment.

The cluster-randomized nature of the design is reflected in analysis at the cluster level, as we propose above.

**TAITU-B Primary Outcome**

The primary outcome is ocular chlamydia infection in the 0-5 year-old age group (individual-level, binary), compared between the targeted azithromycin arm and the delayed mass azithromycin arm in a superiority analysis.

The statistical approach will be to conduct clustered logistic regression, using the baseline values and treatment assignment as covariates. More specifically, the outcome will consist of the vector of 12- and 24-month outcomes, and we will cluster on community (reflecting the randomization) and individual (reflecting the measurement of two observations on the same individual). Note that no adjustment for post-randomization covariates will be performed.

The regression coefficient for the treatment assignment will be used to determine statistical significance. Estimation will be performed using maximum likelihood, using the likelihood ratio test.

Evidence of outliers or strong departure from normality will necessitate the use of robust regression. Permutation based P-values will be reported; all regressions conducted will always be reported. Statistical significance will be determined using a two-sided alpha of 0.05.

**Secondary outcomes**

- Data will be analyzed on several secondary outcomes as described in the first specific aim, but in slightly different age groups (i.e., 0-5 years, 8-12 years):
    - Clinically active trachoma, graded on an ordinal scale
    - Trachoma improvement score comparing baseline vs final visit
    - Serologic evidence of chlamydial antigens from dried blood spots
    - Nasopharyngeal pneumococcal macrolide resistance in 0-5 year-old children
    - Anthropometry

Other secondary outcomes, including clinical trachoma, chlamydial load, ocular chlamydia in different age groups, will be analyzed as described for the WASH trial.

## 3.4 Cost effectiveness Specific Aim

The cost effectiveness analysis consists of two parts: a statistical, trial-based, short term analysis, and a long-term decision analysis based on mathematical modeling. The long term analysis is discussed more fully in the Proposal.

**3.4.1 Short-term, trial-based analysis**

The short term analysis is designed to provide insight into whether each intervention (WASH or targeted antibiotics) is effective for our primary trial outcome of reducing ocular chlamydial infection in children. The time horizon of these analyses will be the duration of each trial (36 months for the WASH trial, 24 months for the targeted antibiotic trial).

For both short term analyses, we will calculate the incremental cost-effectiveness ratio in terms of costs per percent of chlamydia reduction. As a secondary outcome, we will use data from active monitoring of trachoma, soil transmitted helminths, anthropometry, as well as passive monitoring of health clinics to compute an estimate of disability adjusted life years (DALYs) lost. DALY utility values have been recently estimated for several of the conditions we will be measuring (Salomon 2012):

| **Condition** | **DALY utility (95%CI)** |
|---|---|
| Diarrhea: mild | 0.061 (0.036-0.093) |
| Diarrhea: moderate | 0.202 (0.133-0.299) |
| Diarrhea: severe | 0.281 (0.184-0.399) |
| Intestinal nematode infections: symptomatic | 0.030 (0.016-0.048) |
| Infectious disease: acute episode, mild | 0.005 (0.002-0.011) |
| Infectious disease: acute episode, moderate | 0.053 (0.033-0.081) |
| Infectious disease: acute episode, severe | 0.210 (0.139-0.298) |
| Severe wasting | 0.127 (0.081-0.183) |
| Ear pain | 0.018 (0.009-0.031) |

We will estimate the total DALYs lost in each treatment arm by multiplying the DALY value for each condition by the prevalence of the condition by an estimate for the duration of illness (specific to each illness). Note that estimation of DALY values for conditions such as trachoma and stunting has not been specified, and will therefore require us to make assumptions for the outcomes of trachoma or stunting. We will perform numerous sensitivity analyses testing the robustness of our conclusions.

For short term analyses, we propose to report standard cost-effectiveness acceptability curves (e.g. Glick et al, 2007) based on bootstrap resampling at the village level from both control and intervention villages. For control villages, no programmatic costs are incurred, but trachoma outcomes are still available. Plots of resampling estimates in the cost-effectiveness plane will be reported, along with confidence limits for the incremental cost-effectiveness ratio. Note that while it will be possible to compare the WASH intervention and targeted antibiotic intervention at month 12, this comparison will almost certainly favor the targeted antibiotic intervention given the high up-front costs of WASH, and will therefore not be of considerable interest.

**3.4.2 Long-term, decision model-based analysis**

The long term analysis will use data from the results of each trial and extrapolate these results out to a longer time horizon than the trial can offer (e.g., 5 years, 10 years, 15 years, etc.).

**3.4.2.1 Chlamydia effectiveness outcome.** We will use a state transition model using data from Specific Aims 1 and 2 to calculate the effectiveness outcomes for the long-term cost-effectiveness analysis. As in other models of infectious disease control, reduction in community transmission is an important consideration and will be modeled explicitly. We will extend previously developed trachoma models. Specifically, we propose to model the risk of trachoma exposure as an increasing function of the prevalence in the community using a power law, which reduces to a linear function as a special case. For the base case, only infection in children will be considered, although extension to multiple age groups is straightforward. Each community is modeled as a Markov process in continuous time, and Kolmogorov equations for the probability distribution of the number of infected children will be analyzed. Because PCR coverage, although high, is less than 100%, the process is only partially observed. Analysis of the equations yields the likelihood of the observed data at each time point; we will use maximum likelihood methods to fit the transmission model to the data from Specific Aim 1 for the three years of the study (and likewise will model the data from Specific Aim 2 for the 24 months of that study); this process will be repeated for each arm of the study. Standard errors will be computed using bootstrap resampling at the village level. From a model which represents the total number of individuals who are infected and their duration of infection, we can compute the total person-time infected for the analytic horizon (e.g., 10 years), and therefore the expected number of individuals who will eventually be infected. Comparison of the outcomes from the two arms directly estimates the intervention benefit. Note that future health benefits will be discounted.

**3.4.2.2 DALY outcome.** We will create a similar transition model using data from Specific Aims 1 and 2 in order to estimate the DALYs lost in each community.

**3.4.2.3 Decision model.** We will create a decision model with ocular chlamydia as the effectiveness outcome in order to assess the long-term cost-effectiveness of the WASH interventions and targeted antibiotics for trachoma elimination. Using the transition state model described above, we will be able to model different scenarios, including one where trachoma is eliminated by 2020 but WASH interventions (or targeted antibiotics) are needed to maintain elimination. We will calculate the incremental cost-effectiveness ratio as the mean difference in costs between the 2 arms divided by the mean difference in effects, which will be interpreted as the costs per infection-year averted. We will model both costs and effects beyond the horizon of the clinical trial in a simple decision model. The model will have only 1 decision node: (A) the WASH intervention, (B) the targeted azithromycin intervention, (C) mass azithromycin, or (D) no intervention. Costs and effects beyond the trial endpoint (beyond 12 months for the targeted azithromycin trial and beyond 36 months for the WASH trial) will be modeled as described above, and will be discounted at 5%. We will calculate the incremental cost effectiveness at years 5, 10, and 15 after the initiation of the intervention by dividing the average difference in costs by the average difference in effects. At each time point, we will perform probabilistic sensitivity analyses to account for uncertainty, using the distribution of costs and effects observed in the clinical trial. We will display these results in a cost-effectiveness acceptability frontier, which is a graph that shows the probability that a given treatment option is cost-effective over varying levels of willingness to pay, assuming that the treatment option with the

highest net monetary benefit is chosen. We will also construct a decision model using DALYs lost as the effectiveness measure and varying the assumptions about the strength and duration of the effect of each trial's intervention on the morbidity and mortality outcomes.

# 4 Interim monitoring

## 4.1 Data and Safety Monitoring Committee

This trial will be overseen by a Data and Safety Monitoring Board, described elsewhere, who will review procedures, including this Statistical Analysis Plan. The scope and role of the DSMB will be elaborated in the Board's charter.

## 4.2 Interim analysis and stopping rules

No formal interim analysis is proposed for efficacy or futility. All randomization units (communities) are enrolled at the beginning of the trial, and information from some units cannot be used to guide enrollment.

## 4.3 Safety monitoring

Safety monitoring procedures are discussed elsewhere. Note that the WUHA interventions consist of standard hygiene and sanitation interventions (latrines, water wells, and so on as described elsewhere), and the TAITU interventions consist of mass antibiotic distributions that are currently the standard of care recommended by the World Health Organization.

## 4.4 Prevalence monitoring

As discussed in the Protocol and Manual of Operations, ongoing monitoring of trachoma (clinical signs and/or PCR prevalence) may trigger a decision to treat with azithromycin according to WHO guidelines. This trial is not designed to evaluate azithromycin; the retreatment rule is indicated by ethical considerations.

Retreatment occurring at the end of the trial (after our last data collection) at month 36 of course plays no role in the analysis. We propose if retreatment occurs at month 24 (after the 24 month collection) to use collections at 12 month and 24 month as the basis for the analysis, so that the final analysis would proceed as described above simply omitting the 36 month data.

# Appendix

# A Data collection

## A.1 Overall design

Data collection forms, training, security, and quality assurance are discussed in the Manual of Operations. Note that the data security and quality assurance plan includes (1) secure transfer of data from electronic data capture to a database system, (2) regular offsite secure backup systems in case of equipment failure or other unforeseen circumstances, (3) maintenance of confidential patient records using encryption to ensure compliance with applicable regulations and directives, and (4) communication of databases between sites in a secure manner (secure FTP and/or encrypted email). Encryption will be conducted using the Advanced Encryption Standard (AES).

## A.2 Analysis sets

Data sets for analysis will be produced at the Proctor central site by the database manager. Each will be a Microsoft Excel® worksheet containing a single header line whose variable names match the main database. Each analysis set will be in the form of a rectangular table in which each column corresponds to a single variable and each row to an observation. All missing values will be coded explicitly using the string `NA` (as used in the R software). Codes for categorical variables (such as 1 for male, and 2 for female) will be avoided in favor of self-documenting character strings (such as Male, Female) whenever possible. Automated checks will be made to ensure consistency and that each variable in the analysis set has in-range values (protecting against negative ages, spelling errors in categorical factors, and similar errors).

A detailed codebook will be prepared, containing for each variable, (a) the form from which the variable derived, (b) the text of the question (when relevant), (c) all possible values for the variable, and (d) summary statistics for the variable. Note that all codes and character strings that represent categorical factors will be clearly defined in the codebook. Units for each continuous variable will be unambiguously indicated for each variable. Each release of the analysis set will be accompanied by the corresponding version of the codebook. Version numbering with dates will be strictly observed. Standard report-generation software included with the R statistical and data analysis package will be used to ensure consistency of the codebook and analysis set at all times. Backups will be made of all analysis sets, including the queries used to generate them.

## A.3 Data monitoring reports

Data monitoring reports will be prepared based on analysis data sets. These will be prepared using report-generation software. Monitoring reports will include baseline findings as well as intervention reports (i.e., number of installed latrines, date of installation of wells).

All demographic and history variables determined at presentation or enrollment will be summarized by counts and percentages tabulated by community.

Age-specific breakdowns will be available for each community from the baseline census. Additional baseline characteristics may be recorded for each community.

Detailed records of all government health activities (education, vaccination, or treatment programs) in the region will be collected.

# B Power

## B.1 Power for changes in *Chlamydia* load caused by intervention

The planned sample size of 22 communities per arm (44 total communities) provides over 80% power to detect a standardized effect size of 1.07 (effect size divided by the standard deviation), assuming an alpha of 0.05 (two sided) and using the power formula for a two-sample T-test. The standard deviation of the log geometric mean from three villages in Tanzania (Solomon et al, 2003) was 0.5; using this as the standard deviation would imply somewhat more than 80% power to detect an effect size of approximately 0.54 in the log geometric mean.

## B.2 Power for cost-effectiveness analysis

The long-term benefit of the intervention on ocular chlamydia will be assessed using the long-term mathematical model described in the Protocol. This analysis is based on fitting a state-space model to longitudinal data using standard techniques as described in the protocol. In brief, this procedure will be based on a state space transition model for the number of infected children, which will be fitted using maximum likelihood methods to baseline and observed data (and which constitutes, in effect, a nonlinear parametric regression model). The number of infections may be extrapolated forward beyond the end of the study, yielding a time series of infection prevalence data. Discounted future benefits may then be computed from this over the lifetime of the well and latrine interventions, assuming ongoing costs for the hygiene interventions. Bootstrap resampling at the village level will be used to compute standard errors. Note that not only the observed prevalence series for the trial will be used for each resampled village, but the program costs as well; this procedure reflects the correlation between costs and effects seen in the trial.

We assessed the power for the proposed short term, trial-based cost-effectiveness analysis (based on the community-level statistical analysis described in the protocol) using Fieller's method (Formula 9.7, Glick et al, 2007). Analysis of cost data (unpublished) from the Trachoma Amelioration in Amhara study, latrine intervention arm (Stoller et al, 2011), supports use of a conservative standard deviation of 170 USD in the per-village costs. A sample size of 22 communities per arm should provide over 80% power to rule out incremental cost-effectiveness ratios above a willingness to pay threshold of 500 USD, an effect size of 6 person-years of infection averted per community over the course of the study, and a correlation $\rho$ of 1/3. A similar value is found for $\rho=0$. The same value is used for the standard deviation (for prevalence) as in B.1.

It is important to realize that the long-term benefit may be much larger than the short term benefit above, as discussed in the Protocol, and that water, in particular, has many benefits besides trachoma reduction—the value of clean water in promoting health is well-accepted. These additional benefits will be included in sensitivity analysis of the long-term model.

# C Statistical considerations

## C.1 Summarization

Demographic characteristics of participants will be summarized by age and community for each visit.

## C.2 Statistical modeling

For each statistical analysis, model assumptions will be checked. For linear mixed effects modeling, empirical best linear unbiased predictors will be calculated, and plots of residuals versus fitted values will be examined. Leave-one-out influence plots will be examined as well, to ensure that any conclusions are not unduly influenced by any single observation or community. Rank transformed and/or robust linear mixed models (R package `robustlmm`) will be employed if necessary. Outliers are to be expected in assessing chlamydial prevalence data.

For analysis of chlamydial load, sensitivity analysis will be conducted using different community measures of chlamydial load.

We will also conduct exploratory regressions in which individual-level data are modeled using clustered logistic regression. Age will be included as a covariate. Additional community-level analyses will include other community-level covariates, such as altitude or distance to the principal transportation road.

Alternative modeling strategies, including the use of generalized estimating equations and rank based regression, will be conducted to ensure that any results reported were not simply the result of the choice of statistical method. Note that all procedures reflect the clustered nature of the randomization.

In addition to the primary pre-specified analysis, the risk difference, relative hazard ratio, and relative risk will be modeled in generalized linear models with the identity, complementary log-log, and log binomial links, respectively. The results of these analyses will be reported in a way that is statistically congruent with the primary analysis. These alternative values do not constitute separate tests, but rather permit more flexibility in future meta-analyses that may be conducted by readers of our publication.

We emphasize that these analyses are supplementary and will be distinguished from the primary analysis.

For modeling and cost-effectiveness analyses, probabilistic sensitivity analyses will be conducted with respect to all input parameters, including the discount rate.

The study hypotheses are listed as bidirectional effects; we will never conduct one-sided tests.

Effect sizes will be reported along with all hypothesis tests, together with corresponding confidence intervals.

Because the legitimacy of the hypothesis tests being conducted depends on the assumptions (i.e. normality and homoskedasticity for linear models) the adequacy of the statistical model must be checked. Methods which will be employed may include (a) residual plots (vs. baseline value, vs. predicted values, and Q-Q plots), (b) jackknife influence estimates, and (c) when appropriate, tests for normality (including the Anderson-Darling and Shapiro-Wilk procedures). In every case, regression and modeling assumptions will be checked and diagnostics reported.

## C.3 Missing data

The intent-to-treat principle governs our analysis. We will report the results of complete case analysis when there are missing communities; community-level indicators will be based on available 0-5 year olds. Sensitivity analysis based on assigning values to unobserved data and computing the statistical analysis will be conducted to assess the degree to which missing data could have affected our conclusions and will always be clearly identified as such and separated from the main analysis. Additionally, regression-based multiple imputation will be conducted based on the assumption of missingness at random.

## C.4 Software

The standard software package R (`http://www.r-project.org`) for the MacIntosh OS X will be used for all descriptive and inferential analyses. Mixed modeling will be conducted using the `lme4` package for R.

## C.5 Data management

Data management is discussed in the Manual of Operations. The data management plan is designed to ensure information security (prevention of unauthorized access and maintenance of confidentiality) and information integrity (logging of all changes, maintenance of backups on and offsite).

# D Randomization procedures

Communities who are eligible for WUHA because they contain a potential water point in a unique school district will be allocated at random to two groups: 1) intervention, and 2) control. For TAITU, communities from the study region will be allocated at random to three groups: 1) targeted treatment, 3) MDA and 2) control. These randomizations will be done *subsequent to* the baseline census and monitoring visit. For definiteness, the procedure will be as follows:

- We will use a digital pseudorandom number generator for which a random number seed will be identified. The random number seed is completely arbitrary and has no significance other than its use to create a random assignment. Choice of the seed completely determines the assignment. Identification of the random number seed is conducted with two goals: (1) The seed cannot be known by masked personnel, and (2) the seed must yield assignments which cannot be influenced by study personnel. Thus, (a) the seed cannot be known to masked personnel during the study, and (b) the seed must be determined from values which obviously are beyond the control of study personnel. This is accomplished by prespecifing the use of numbers which are not known at the time of design and which manifestly cannot be influenced by study personnel; future meteorological measurements in specified cities, stock market indices for future dates, or other similar values are all acceptable choices.
- The seed will be used to initialize the random number generator for the R statistics package (using the default algorithm).
- A list of all communities will be alphabetized and entered as a character vector in the R package.
- The `sample` function will be called without replacement on this list yielding a shuffled version. This call is to be the first use of the random number generator following seeding.
- Communities will be read off the shuffled list in sequential order until the sample size for the intervention group has been achieved, and those communities are then randomized into the intervention arm of the study.

The randomization lists will be prepared by the Proctor center and communicated to the site representative in Ethiopia. TP will print a hard copy of the randomization list and keep it stored in a locked file cabinet after deleting the electronic copy from his computer and email account.

Distribution of the randomization lists to Ethiopia will be accomplished using the University of California, San Francisco's encrypted email provision. Email is encrypted using the Advanced Encryption Standard (NIST FIPS 197) whenever the first four characters of the subject line are `PHI:`. The sender is notified when the recipient receives a secure email; the recipient receives a notification of a secure email and can view it using the UCSF Secure Messenger website. We have successfully used this method in previous studies.

A backup copy of the full randomization lists will be maintained by Proctor Director Tom Lietman, MD. These lists will be maintained as a hard copies stored in a locked file cabinet at the Proctor site.

As discussed below, the randomization lists will be provided as Excel® worksheets. No technical knowledge will be required to use these lists.

# E Reporting conventions

- All tables and data listings will be presented in landscape orientation, unless presented as part of the text of the final report.
- Figures will be presented in landscape orientation, unless the information is substantially easier to interpret in portrait orientation.
- Direct annotation of figures will be preferred to legends. All figures with more than one variable or item will contain either direct annotation or legends. All annotation will be unambiguously identifiable as such.
- Color will be used in figures only when needed to enhance clarity of communication. All color schemes will be evaluated for visual clarity for individuals with diminished color vision. All color encodings will be identified. Redundant encodings (such as the use of different plot symbols or line dash patterns) will be used in addition to color, so that all figures are interpretable after monochrome reproduction at 100 dots per inch. All dash patterns and line widths will be adequate to be distinguishable after monochrome reproduction at 100 dots per inch. Any distinction between plot symbols (circles, filled circles, diamonds, etc.) will remain clear after monochrome reproduction at 100 dots per inch.
- Fixed width sans serif fonts will be used for all labeling (Helvetica, Arial, or Futura).
- Boldface and italics will not be used unless substantial value is added.
- Decorative fonts and enhancements, including borders and shading, will not be used. Decorative presentation methods, such as ribbon graphs, will never be used.
- All information given in figures will also be presented in summary tables (perhaps only included in an Appendix or in supplementary materials).
- Only standard characters will be used in tables and data listings.
- All titles will be centered. The first title line will be the number of the table, figure, or listing. The second and possibly third lines will be the description of the table, figure, or data listing. The ICH numbering convention will be used for all.
- All footnotes will be left justified and at the page bottom. Footnotes will be used sparingly. Reference footnotes will be complete enough to locate any reference based on the information provided (Author, Journal, Pages, Date, or PubMed accession number).
- Missing values for numeric or character variables will be unambiguously identified as such using the special string `NA` (not available) in all settings; `NA` is the standard missing value code for our software. Each figure or table caption in which `NA` is used will indicate the meaning of `NA` in that figure or table. The abbreviation `NA` will never be used for any other purpose.
- All date values will presented in the form DDmmmYYYY format (e.g. 01jan2008), using four digit years. June will be encoded as `jne` (otherwise `jan` and `jun` would differ by only a single character), and July as `jly` (so that the lowercase letter `l`, easily confused with the digit `1`, will not be adjacent to any numerals).
- All tables, figures, and data listings will have the name of the program and a date/time stamp on the bottom of the output.

# F Revisions

- September 26, 2017: Added updated sample size calculations for WUHA
- November 11, 2019: Added additional secondary analysis techniques to Appendix C; updated sample size calculation and analysis plan for Specific Aim 2B
- October 29, 2020: Specified that time assessed as continuous variable in primary WUHA analysis

# References

Chow SC, Shao J, Wang H. Sample Size Calculations in Clinical Research, Second Edition, Chapman & Hall/CRC, Boca Raton, 2007.

Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic evaluation in clinical trials. Oxford University Press, Oxford, 2007.

Solomon AW, Holland MJ, Burton MJ, West SK, Alexander NDE, Aguirre A, Massae PA, Mkocha H, Muñoz B, Johnson GJ, Peeling RW, Bailey RL, Foster A, Mabey DCW. Strategies for control of trachoma: observational study with quantitative PCR. *Lancet* 262:198-204, 2003.

Salomon, JA, Vos, T, Hogan, DR, Gagnon, M, Naghavi, M, Mokdad, A, et al. Common values in assessing health outcomes from disease and injury: diability weights measurement study for the Global Burden of Disease Study 2010. Lancet 2012; 380:2129-43.

# Sanitation, Water, and Instruction in Face-washing for Trachoma II

# Statistical Analysis Plan

Confidential

Version 1.1, 29 October 2020

Francis I. Proctor Foundation for Research in Ophthalmology

# 1 Introduction

This document (Statistical Analysis Plan, SAP) describes the planned analysis and reporting for the **Sanitation, Water, and Instruction in Face-Washing Trial II (SWIFT II),** University of California, San Francisco, J. Keenan, PI. It includes specifications for the statistical analyses and tables to be prepared for the final Clinical Study Reports.

The original SWIFT project consisted of two cluster-randomized trials: WUHA and TAITU. In SWIFT II, we continue study of the WUHA study clusters only. During WUHA I, 20 study clusters were randomized to a water, sanitation, and hygiene (WASH) intervention and 20 study clusters to no WASH intervention. In WUHA II, annual mass azithromycin treatments will be distributed in all 40 study clusters, and the 20 clusters randomized to WASH during WUHA I will continue to receive WASH while the 20 clusters randomized to control will continue to not receive the WASH intervention.

The content of this Statistical Analysis Plan meets the requirements stated by the US Food and Drug Administration (Department of Health and Human Services, Food and Drug Administration, 1998) and conforms to the American Statistical Association's Ethical Guidelines (American Statistical Association, 1999).

The following documents were reviewed in preparation of this Statistical Analysis Plan:

- Trial Manual of Operations, and Proposal
- ICH Guidance on Statistical Principles for Clinical Trials (Department of Health and Human Services, Food and Drug Administration, 1998)
- Statistical Analysis Plan (prepared by T. Porco), Trachoma Amelioration in Amhara II (T. Lietman, PI)

The planned analyses described in this Statistical Analysis Plan will be included in future manuscripts. Note, however, that exploratory analyses not necessarily identified in this Statistical Analysis Plan may be performed to support the analysis. All post-hoc or unplanned analyses which have not been delineated in this Statistical Analysis Plan will be clearly documented as such in the final Clinical Study Report, manuscripts, or any other document or submission.

Finalization of this document will take place prior to the final study visit; the final version of this document will be numbered 2.0. All subsequent changes will be indicated by detailed change log in an Appendix.

J. Keenan, D. Fry, T. Lietman, and T. Porco have contributed to this plan.

# 2 Design

The study consists of a cluster randomized trial, designated as the "WUHA II" trial." Alongside the trial, multiple tests for ocular chlamydia will be collected in order to determine the diagnostic accuracy of each. The design of the trials is described fully in the Proposal and the Manual of Operations. Masking, selection of communities, inclusion criteria, exclusion criteria, and informed consent are discussed in those documents.

## 2.1 Outcomes

The design is community-randomized, based on school districts. Details are provided in the Protocol (WUHA Research Plan).

Interventions and monitoring will take place only in a subset of individuals within each school district (i.e., the study cluster), as indicated in the Protocol. The primary outcome variable is the prevalence of *Chlamydia trachomatis* in children, as assessed from PCR of conjunctival swabs. The age ranges, sampling frame, and grading or laboratory methods are described in the Protocol.

## 2.2 Sample size

The sample size for the WASH trial is based on the primary outcome of Specific Aim 1: the cluster-level prevalence of ocular chlamydia. Power is reported for secondary aims and for the other Specific Aims (Appendix).

The sample size for WUHA II is fixed, based on the sample size of WUHA I. Sample size calculations for WUHA II assume a repeated measures design, with outcomes assessed at four time points (month 48, 60, 72, and 84). We assume that the prevalence of chlamydia has a correlation of 0.45 between study visits (based on data from the month 0 and month 12 visits of WUHA I). If we assume a standard deviation of 0.12 (from WUHA I) and a two-sided alpha of 0.05, then 20 communities per arm will provide 80% power to detect an effect size of 0.081 (8.1%). Prior experience of TEF, TANA, TIRET, and WUHA supports the assumption that no communities will be lost to follow-up. The sample size plan is based on a cluster level analysis and therefore reflects the clustered nature of the design.

The primary analysis for Specific Aim 2 is a hidden Markov model, using individual-level data from the 4 time points. Sample size calculation is not straightforward. The major goal of Specific Aim 2 is to estimate the sensitivity and specificity of several different tests for ocular chlamydia. After the mass azithromycin treatment, we assume the prevalence of ocular chlamydia will be brought to a low level, but some infections will persist. As a lower estimate of the power of this analysis, if we assume that 5% of children are infected at each visit, this would yield 120 positive tests out of 2400 children. Thus, if a test has a true sensitivity and specificity of 95%, then we would be able to estimate that sensitivity with a total 95% confidence interval width within 9%

and specificity within 2%. In reality, estimates of diagnostic accuracy tests should have even more precision since we include multiple time points.

### 2.3 Randomization

Procedures for site selection, community eligibility criteria, and informed consent procedures are discussed in the Protocol and the Manual of Operations. WUHA I has already enrolled 40 communities and these same communities will be analyzed in their original randomization groups in WUHA II.

# 3 Primary hypothesis tests

Procedures for handling missing data, checking model assumptions, and conducting secondary analyses are provided in the Appendix.

## 3.1 Specific Aim 1

**Primary outcome** . The following information is anticipated:

1. $Y_{it}$, the prevalence of infection in village $i$ at time $t$, where $t$=1,2, or 3 (year 1, 2, or 3). This is the outcome variable.
2. $P_i$, the baseline prevalence in village $i$ (which could have been denoted $Y_{i0}$). As discussed elsewhere, this is the prevalence in children aged 0-5 as determined by PCR testing. The age limit of 5 is inclusive. Specifically, this is the baseline at time t=0, before randomization.
3. $X_i$, the treatment covariate for each village, 0 for control group, and 1 for intervention group.
4. $T_{it}$, the time the infection data was collected in village $i$ at time $t$, in continuous months after baseline.

We wish to use the four post-antibiotic time points as part of the outcome using a pooled scheme. Several possibilities must be considered: (a) the effect of the antibiotic intervention is fast, and the village-specific prevalences in the treatment arms are lower than baseline, (b) the village-specific prevalences will be lower in the intervention than the control groups, and show no further trend: the slopes are the same, but the mean is lower in the treatment group than the control group, and (b) the effect of the intervention is gradual, with the difference between the control and intervention group becoming larger over time.

We propose to fit a linear mixed effects regression model to the village-specific prevalences over time:

$$Y_{it} = \beta_0 + \beta_1 P_i + \beta_2 T_{it} + \beta_3 X_i$$

We will conduct several sensitivity analyses. First we will repeat the same pooled regression analysis but include data from all time points since the original randomization (i.e., months 12, 24, 36, 48, 60, 72, and 84). We note that the period of time from baseline to month 36 is fundamentally different from the period of time from month 48 to 84, and will explore the use of an indicator variable for these two time periods.

In a second set of sensitivity analyses, we will perform a similar mixed effects linear regression but also include a random slope and intercept term because successive measurements on the same village may show statistical dependence. For this latter sensitivity analysis the fixed effects will be modeled according to the following equation:

$$Y_{it} = \beta_0 + \beta_1 P_i + \beta_2 T_{it} + \beta_3 X_i + \beta_4 T_{it} X_i$$

We propose to include the interaction term only if it is significant. We will perform the analysis using data from all time points from the study and will perform a separate analysis using only the time points from WUHA II.

If community participation becomes very low (e.g., <50%) in some communities, inverse probability weighting will be considered to ensure that one anomalous value from a small sample does not predominate.

For all analyses, estimation will be conducted by restricted maximum likelihood. Testing will be two sided at an alpha level of 0.05, based on the value of $\beta_3$, testing the difference between the control and intervention communities. If the interaction term is included, then the deviance statistic contrasting the model with all terms with the model without $\beta_3$ and $\beta_4$ will be used.

Significance testing will be conducted by permutation testing at the village level, and will reflect the entire fitting algorithm, preserving the overall alpha level at 0.05. We anticipate modestly higher power than used for planning, since data from all three follow-up visits will be used, and not just the data at the final visit.

Descriptive statistics will be reported for each wave of study.

Diagnostics, alternatives, and sensitivity analyses are discussed in the Appendix.

## 3.2 Additional Secondary analyses, Specific Aim 1

Diagnostics, alternatives, and sensitivity analyses are discussed in the Appendix.

3.2.1. Village-level *Chlamydia* load outcome.
The analysis is proposed to be identical to that of the ocular chlamydia outcome in §3.1, except that a village-specific index of chlamydia load at baseline and at follow-up times is used instead of prevalence. The analysis is two-sided at an alpha of 0.05.

This is a prespecified secondary analysis and will be reported as such.

3.2.2. Follicular trachoma scores, longitudinal analysis.
Follicular trachoma scores (using the 5-level system) will be modeled longitudinally using linear mixed effects regression. Scores are indexed by participant, grader, visit, and village. Participant and village will be modeled as random effects; grader will be modeled as a fixed effect. We will use an AR(1) correlation structure.

3.2.3. Inflammatory trachoma scores, longitudinal analysis.
Inflammatory trachoma grades will be modeled in the same way as in Section 3.2.2.

3.2.4. Chlamydial load, individual level analysis
Quantitative PCR results at the individual level will be modeled using standard procedures for semi-continuous variables.

3.2.5 Ocular chlamydia, age 6-9 or older
Longitudinal analysis of ocular chlamydia in the age 6-9 group will be modeled at the village level. A similar analysis will consider the 10 and over segment of the population.

3.2.6 Nasopharyngeal macrolide resistance
Nasopharyngeal macrolide resistance in age 0-5 will be modeled at the village level, using treatment arm as a covariate.

3.2.7 Anthropometry
Longitudinal analysis of anthropometric measurements between the two arms will be conducted using growth curve models.

3.2.8 Clinical trachoma
We anticipate having the following information available. For each child in the sampling frame, we will have a binary improvement score, based on baseline and follow-up photography. We propose to conduct clustered logistic regression (taking into account the clustered nature of the design) using village assignment as the predictor. We will estimate the log odds of the treatment effect. Significance testing will be conducted at 0.05 based on Monte Carlo permutation testing.

3.2.9. Prevalence of chlamydia antigen positivity from serological tests (village-level analysis)
The prevalence of soil transmitted helminths will be compared between the treatment arms in a linear mixed effects regression as described above for ocular chlamydia.

## 3.3 Specific Aim 2

**Primary Analysis.** The primary analysis uses a hidden Markov model (HMM) to assess the sensitivity and specificity of several diagnostic tests for assessing a true state of infection. This approach does not require a gold standard diagnostic test, but instead considers an unobserved latent, or hidden, health state (e.g., true infection in an individual, or true prevalence in a community). In our case, the HMM will consist of a structural model for ocular chlamydial infection and a measurement model for the observed diagnostic tests, with the diagnostic tests conditional on the latent health state. The following diagnostic tests will be included, which fall into 3 classes of tests:

- Tests of clinically active trachoma
    - TF and TI from field grading (i.e., grades according to the World Health Organization simplified trachoma grading scale)
    - Follicular and Inflammatory scores as graded from photographic review of conjunctival photographs
    - Automated algorithm: presence of TF and TI from a convolutional neural network classifier
- Tests of ocular chlamydia
    - Abbott m2000 RealTime chlamydia PCR
    - Biomeme quantitative PCR
    - miniPCR
    - Loop-mediated isothermal amplification (LAMP)
- Serologic tests
    - pgp3 seropositivity, assessed from Luminex platform
    - CT694 seropositivity, assessed from Luminex platform
    - pgp3 seropositivity, assessed from a lateral flow assay (LFA)

The structure of the HMM can account for hierarchical clustering (e.g., children within households within communities) as well as information from prior health states (e.g., chlamydial infection 12 months prior) to inform the probability of infection at the present visit. This approach is well suited for trachoma, since some diagnostic tests for trachoma are more indicative of prior infection (e.g., TF, serology) whereas others suggest current infection (PCR). Estimates of sensitivity, specificity, and positive and negative predictive value, as well as confidence intervals around differences between these metrics between pairs of tests, will be estimated using the standard Metropolis-Hastings algorithm. We propose the use of the R packages `mcmc` and `coda` (though others are serviceable), with chain convergence assessed using the Gelman-Rubin statistic. Our group has prior experience with inference based on Markov chain Monte Carlo.(Deiner 2017; Lietman 2011).

**Supplementary Analysis.** The analysis above models individual-level data. We will also construct a model that models the prevalence of test positivity from each community over time, using the methods we have applied previously (Liu, 2015).

**Secondary outcomes**

- Abbott gold standard: we will calculate the sensitivity and specificity of the PCR tests at

the individual level assuming the Abbott RealTime assay as the reference standard.

- Latent class analysis of individual nucleic acid amplification tests (NAATs): We acknowledge that the Abbott RealTime assay is not a true gold standard test. Therefore, we will also perform Latent Class Analysis (LCA) at cross-sectional time points.

## 3.4 Cost effectiveness Analysis

The cost effectiveness analysis consists of two parts: a statistical, trial-based, short term analysis, and a long-term decision analysis based on mathematical modeling. The long-term analysis is discussed more fully in the Proposal.

### 3.4.1 Short-term, trial-based analysis

The short-term analysis is designed to provide insight into whether WASH is effective for our primary trial outcome of reducing ocular chlamydial infection in children. The time horizon of these analyses will be the duration of the trial (48 months).

We will calculate the incremental cost-effectiveness ratio in terms of costs per percent of chlamydia infection prevented. We propose to report standard cost-effectiveness acceptability curves (e.g. Glick et al, 2007) based on bootstrap resampling at the village level from both control and intervention villages. For control villages, no programmatic costs are incurred, but trachoma outcomes are still available. Plots of resampling estimates in the cost-effectiveness plane will be reported, along with confidence limits for the incremental cost-effectiveness ratio.

### 3.4.2 Long-term, decision model-based analysis

The long-term analysis will use data from the results of each trial and extrapolate these results out to a longer time horizon than the trial can offer (e.g., 5 years, 10 years, 15 years, etc.).

**3.4.2.1 Chlamydia effectiveness outcome.** We will use a state transition model using data from Specific Aims 1 and 2 to calculate the effectiveness outcomes for the long-term cost-effectiveness analysis. As in other models of infectious disease control, reduction in community transmission is an important consideration and will be modeled explicitly. We will extend previously developed trachoma models. Specifically, we propose to model the risk of trachoma exposure as an increasing function of the prevalence in the community using a power law, which reduces to a linear function as a special case. For the base case, only infection in children will be considered, although extension to multiple age groups is straightforward. Each community is modeled as a Markov process in continuous time, and Kolmogorov equations for the probability distribution of the number of infected children will be analyzed. Because PCR coverage, although high, is less than 100%, the process is only partially observed. Analysis of the equations yields the likelihood of the observed data at each time point; we will use maximum likelihood methods to fit the transmission model to the data from Specific Aim 1 for the four years of the study; this process will be repeated for each arm of the study. Standard errors will be computed using bootstrap resampling at the community level. From a model which represents the total number of individuals who are infected and their duration of infection, we can compute the total person-time infected for the analytic horizon (e.g., 10 years), and therefore the expected number

of individuals who will eventually be infected. Comparison of the outcomes from the two arms directly estimates the intervention benefit. Note that future health benefits will be discounted.

**3.4.2.3 Decision model.** We will create a decision model with ocular chlamydia as the effectiveness outcome in order to assess the long-term cost-effectiveness of the WASH interventions for trachoma elimination. Using the transition state model described above, we will be able to model different scenarios, including one where trachoma is eliminated but WASH interventions are needed to maintain elimination. We will calculate the incremental cost-effectiveness ratio as the mean difference in costs between the 2 arms divided by the mean difference in effects, which will be interpreted as the costs per infection-year averted. We will model both costs and effects beyond the horizon of the clinical trial in a simple decision model. The model will have only 1 decision node: (A) the WASH intervention or (B) no intervention. Costs and effects beyond the trial endpoint will be modeled as described above, and will be discounted at 5%. We will calculate the incremental cost effectiveness at years 5, 10, and 15 after the initiation of the intervention by dividing the average difference in costs by the average difference in effects. At each time point, we will perform probabilistic sensitivity analyses to account for uncertainty, using the distribution of costs and effects observed in the clinical trial. We will display these results in a cost-effectiveness acceptability frontier, which is a graph that shows the probability that a given treatment option is cost-effective over varying levels of willingness to pay, assuming that the treatment option with the highest net monetary benefit is chosen.

# 4 Interim monitoring

## 4.1 Data and Safety Monitoring Committee

This trial will be overseen by a Data and Safety Monitoring Committee (DSMC), described elsewhere, who will review procedures, including this Statistical Analysis Plan. The scope and role of the DSMB will be elaborated in the Board's charter.

## 4.2 Interim analysis and stopping rules

The following interim analysis plan will be presented to and must be approved by the DSMC before the start of the continuation study. A formal interim analysis is proposed for efficacy and futility after the mid-point of the study (i.e., after the chlamydia data is complete for month 60) using the same statistical models to be used for the final analysis. All randomization units (communities) are enrolled at the beginning of the trial, and information from some units cannot be used to guide enrollment.

### 4.2.1 Efficacy

Unmasked interim analyses will be conducted to determine whether or not sufficient evidence has accumulated to justify stopping the trial because one treatment is clearly superior (and therefore should be extended to all future cases). The guidelines for efficacy will use group sequential boundaries for judging the statistical significance of the primary outcome measure (month-60 PCR). The Lan and DeMets flexible alpha spending approach will be used with a power use function such that the two-sided P-value to stop the trial for efficacy is 0.001. The use of a flexible alpha-spending function protects the 0.05 alpha level of the overall trial while allowing for additional interim analyses for efficacy if needed, without specifying the number and timing of the analyses at the start of the study. We note that the alpha spending function cannot be changed once the continuation trial has begun.

### 4.2.2 Futility

Early discontinuation due to the unlikeliness of significant findings conditional on interim results may be considered, based on the original sample size considerations. For evaluating futility, we propose using the B-value approach of Lan and Wittes to calculate conditional power of the study given the observed data. Such power calculations will be performed for a range of possible alternatives including the observed treatment difference in the data as well as the original alternative hypothesis. We propose to stop the study if conditional power for a 8% benefit of the WASH intervention is less than 0.2. Since stopping the trial for futility does not reject the null hypothesis, there is no inflation of size associated with early stopping. Instead, it modestly decreases study power.

### 4.2.3 Execution of interim analysis

The principal statistician (TP) will conduct the interim analysis in an unmasked manner, subject to independent statistical review by the DSMC. Quality assurance will be conducted by the database manager.

## 4.3 Safety monitoring

Safety monitoring procedures are discussed elsewhere. Note that the interventions consist of standard hygiene and sanitation interventions (latrines, water wells, and so on as described elsewhere) as well as approved antibiotic treatments.

## 4.4 Prevalence monitoring

As discussed in the Protocol and Manual of Operations, ongoing monitoring of trachoma (clinical signs and/or PCR prevalence) may trigger a decision to treat with azithromycin according to WHO guidelines. This trial is not designed to evaluate azithromycin; the retreatment rule is indicated by ethical considerations.

Retreatment occurring at the end of the trial (after our last data collection) at month 84 of course plays no role in the analysis. We propose if retreatment occurs prior to month 84 to use the most recent pre-retreatment exam visit as the basis for the analysis.

# Appendix

# A Data collection

## A.1 Overall design

Data collection forms, training, security, and quality assurance are discussed in the Manual of Operations. Note that the data security and quality assurance plan includes (1) secure transfer of data from electronic data capture to a database system, (2) regular offsite secure backup systems in case of equipment failure or other unforeseen circumstances, (3) maintenance of confidential patient records using encryption to ensure compliance with applicable regulations and directives, and (4) communication of databases between sites in a secure manner (secure FTP and/or encrypted email). Encryption will be conducted using the Advanced Encryption Standard (AES).

## A.2 Analysis sets

Data sets for analysis will be produced at the Proctor central site by the database manager. Each will be a Microsoft Excel® worksheet containing a single header line whose variable names match the main database. Each analysis set will be in the form of a rectangular table in which each column corresponds to a single variable and each row to an observation. All missing values will be coded explicitly using the string `NA` (as used in the R software). Codes for categorical variables (such as 1 for male, and 2 for female) will be avoided in favor of self-documenting character strings (such as Male, Female) whenever possible. Automated checks will be made to ensure consistency and that each variable in the analysis set has in-range values (protecting against negative ages, spelling errors in categorical factors, and similar errors).

A detailed codebook will be prepared, containing for each variable, (a) the form from which the variable derived, (b) the text of the question (when relevant), (c) all possible values for the variable, and (d) summary statistics for the variable. Note that all codes and character strings that represent categorical factors will be clearly defined in the codebook. Units for each continuous variable will be unambiguously indicated for each variable. Each release of the analysis set will be accompanied by the corresponding version of the codebook. Version numbering with dates will be strictly observed. Standard report-generation software included with the R statistical and data analysis package will be used to ensure consistency of the codebook and analysis set at all times. Backups will be made of all analysis sets, including the queries used to generate them.

## A.3 Data monitoring reports

Data monitoring reports will be prepared based on analysis data sets. These will be prepared using report-generation software. Monitoring reports will include baseline findings as well as intervention reports (i.e., number of installed latrines, date of installation of wells).

All demographic and history variables determined at presentation or enrollment will be summarized by counts and percentages tabulated by community.

Age-specific breakdowns will be available for each community from the baseline census. Additional baseline characteristics may be recorded for each community.

Detailed records of all government health activities (education, vaccination, or treatment programs) in the region will be collected.

# B Power

## B.1 Power for secondary outcomes

The power for each of the secondary outcomes is estimated for the final study visit. In reality, each comparison should have more power since data from 4 study visits will be included.

### Chlamydia load

The planned sample size of 20 communities per arm (40 total communities) provides over 80% power to detect a 33% difference in the log-transformed chlamydial load, assuming a mean log-transformed load of 1.8, a standard deviation of 0.75, an intraclass correlation coefficient (between communities) of 0.121, 10 children with a positive chlamydia test per community, and a two-sided alpha of 0.05. Assumptions are based on individual-level preliminary data from WUHA I.

### Clinically active trachoma

The planned sample size of 20 communities per arm (40 total communities) provides over 80% power to detect a 17% difference in the prevalence of follicular trachoma (TF) between the study arms, assuming a prevalence of 40% in control communities, a standard deviation of 18%, and a two-sided alpha of 0.05. Assumptions are based on community-level preliminary data from WUHA I.

### Chlamydial serology

The planned sample size of 20 communities per arm (40 total communities) provides over 80% power to detect a 17% difference in the prevalence of pgp3 seropositivity between the study arms, assuming a prevalence of 47% in control communities, a standard deviation of 18%, and a two-sided alpha of 0.05. Assumptions are based on community-level preliminary data from WUHA I.

### Nasopharyngeal macrolide resistance

The planned sample size of 20 communities per arm (40 total communities) provides over 80% power to detect a 10% difference in the prevalence of nasopharyngeal pneumococcal macrolide resistance between the study arms, assuming a prevalence of 10% in control communities, a standard deviation of 10%, and a two-sided alpha of 0.05. Assumptions are based on community-level data from the TANA study.

### Anthropometry

The planned sample size of 20 communities per arm (40 total communities) provides over 80% power to detect a 1cm difference in the log-transformed chlamydial load, assuming an average height of 90cm, a standard deviation of 1ocm, an intraclass correlation coefficient (between communities) of 0.2, 30 children per community, and a two-sided alpha of 0.05. Assumptions are based on individual-level preliminary data from the TANA II trial.

## B.2 Power for cost-effectiveness analysis

The long-term benefit of the intervention on ocular chlamydia will be assessed using the long-term mathematical model described in the Protocol. This analysis is based on fitting a state-space model to longitudinal data using standard techniques as described in the protocol. In brief, this procedure will be based on a state space transition model for the number of infected children, which will be fitted using maximum likelihood methods to baseline and observed data (and which constitutes, in effect, a nonlinear parametric regression model). The number of infections may be extrapolated forward beyond the end of the study, yielding a time series of infection prevalence data. Discounted future benefits may then be computed from this over the lifetime of the well and latrine interventions, assuming ongoing costs for the hygiene interventions. Bootstrap resampling at the village level will be used to compute standard errors. Note that not only the observed prevalence series for the trial will be used for each resampled village, but the program costs as well; this procedure reflects the correlation between costs and effects seen in the trial.

We assessed the power for the proposed short term, trial-based cost-effectiveness analysis (based on the community-level statistical analysis described in the protocol) using Fieller's method (Formula 9.7, Glick et al, 2007). Analysis of cost data (unpublished) from the Trachoma Amelioration in Amhara study, latrine intervention arm (Stoller et al, 2011), supports use of a conservative standard deviation of 170 USD in the per-village costs. A sample size of 22 communities per arm should provide over 80% power to rule out incremental cost-effectiveness ratios above a willingness to pay threshold of 500 USD, an effect size of 6 person-years of infection averted per community over the course of the study, and a correlation $\rho$ of 1/3. A similar value is found for $\rho=0$. The same value is used for the standard deviation (for prevalence) as in B.1.

It is important to realize that the long-term benefit may be much larger than the short-term benefit above, as discussed in the Protocol, and that water, in particular, has many benefits besides trachoma reduction—the value of clean water in promoting health is well-accepted. These additional benefits will be included in sensitivity analysis of the long-term model.

# C Statistical considerations

## C.1 Summarization

Demographic characteristics of participants will be summarized by age and community for each visit.

## C.2 Statistical modeling

For each statistical analysis, model assumptions will be checked. For linear mixed effects modeling, empirical best linear unbiased predictors will be calculated, and plots of residuals versus fitted values will be examined. Leave-one-out influence plots will be examined as well, to ensure that any conclusions are not unduly influenced by any single observation or community. Rank transformed and/or robust linear mixed models (R package `robustlmm`) will be employed if necessary. Outliers are to be expected in assessing chlamydial prevalence data.

For analysis of chlamydial load, sensitivity analysis will be conducted using different community measures of chlamydial load.

We will also conduct exploratory regressions in which individual-level data are modeled using clustered logistic regression. Age will be included as a covariate. Additional community-level analyses will include other community-level covariates, such as altitude or distance to the principal transportation road.

Alternative modeling strategies, including the use of generalized estimating equations and rank based regression, will be conducted to ensure that any results reported were not simply the result of the choice of statistical method. Note that all procedures reflect the clustered nature of the randomization.

We emphasize that these analyses are supplementary and will be distinguished from the primary analysis.

For modeling and cost-effectiveness analyses, probabilistic sensitivity analyses will be conducted with respect to all input parameters, including the discount rate.

The study hypotheses are listed as bidirectional effects; we will never conduct one-sided tests.

Effect sizes will be reported along with all hypothesis tests, together with corresponding confidence intervals.

Because the legitimacy of the hypothesis tests being conducted depends on the assumptions (i.e. normality and homoskedasticity for linear models) the adequacy of the statistical model must be checked. Methods which will be employed may include (a) residual plots (vs. baseline value, vs. predicted values, and Q-Q plots), (b) jackknife influence estimates, and (c) when appropriate, tests for normality (including the Anderson-Darling and Shapiro-Wilk procedures). In every case, regression and modeling assumptions will be checked and diagnostics reported.

## C.3 Missing data

The intent-to-treat principle governs our analysis. We will report the results of complete case analysis when there are missing communities; community-level indicators will be based on available 0-5 year olds. Sensitivity analysis based on assigning values to unobserved data and computing the statistical analysis will be conducted to assess the degree to which missing data could have affected our conclusions and will always be clearly identified as such and separated from the main analysis. Additionally, regression-based multiple imputation will be conducted based on the assumption of missingness at random.

## C.4 Software

The standard software package R (`http://www.r-project.org`) for the MacIntosh OS X will be used for all descriptive and inferential analyses. Mixed modeling will be conducted using the `lme4` package for R.

## C.5 Data management

Data management is discussed in the Manual of Operations. The data management plan is designed to ensure information security (prevention of unauthorized access and maintenance of confidentiality) and information integrity (logging of all changes, maintenance of backups on and offsite).

# D Randomization procedures

Communities have already been randomized as part of the WUHA I trial. The following section describes the randomization procedure for WUHA I.

Communities eligible for WUHA because they contain a potential water point in a unique school district will be allocated at random to two groups: 1) intervention, and 2) control. These randomizations will be done *subsequent to* the baseline census and monitoring visit. For definiteness, the procedure will be as follows:

- We will use a digital pseudorandom number generator for which a random number seed will be identified. The random number seed is completely arbitrary and has no significance other than its use to create a random assignment. Choice of the seed completely determines the assignment. Identification of the random number seed is conducted with two goals: (1) The seed cannot be known by masked personnel, and (2) the seed must yield assignments which cannot be influenced by study personnel. Thus, (a) the seed cannot be known to masked personnel during the study, and (b) the seed must be determined from values which obviously are beyond the control of study personnel. This is accomplished by prespecifing the use of numbers which are not known at the time of design and which manifestly cannot be influenced by study personnel; future meteorological measurements in specified cities, stock market indices for future dates, or other similar values are all acceptable choices.
- The seed will be used to initialize the random number generator for the R statistics package (using the default algorithm).
- A list of all communities will be alphabetized and entered as a character vector in the R package.
- The `sample` function will be called without replacement on this list yielding a shuffled version. This call is to be the first use of the random number generator following seeding.
- Communities will be read off the shuffled list in sequential order until the sample size for the intervention group has been achieved, and those communities are then randomized into the intervention arm of the study.

The randomization lists will be prepared by the Proctor center and communicated to the site representative in Ethiopia. TP will print a hard copy of the randomization list and keep it stored in a locked file cabinet after deleting the electronic copy from his computer and email account.

Distribution of the randomization lists to Ethiopia will be accomplished using the University of California, San Francisco's encrypted email provision. Email is encrypted using the Advanced Encryption Standard (NIST FIPS 197) whenever the first four characters of the subject line are `PHI:`. The sender is notified when the recipient receives a secure email; the recipient receives a notification of a secure email and can view it using the UCSF Secure Messenger website. We have successfully used this method in previous studies.

A backup copy of the full randomization lists will be maintained by Proctor Director Tom Lietman, MD. These lists will be maintained as a hard copies stored in a locked file cabinet at the Proctor site.

As discussed below, the randomization lists will be provided as Excel® worksheets. No technical knowledge will be required to use these lists.

# E Reporting conventions

- All tables and data listings will be presented in landscape orientation, unless presented as part of the text of the final report.
- Figures will be presented in landscape orientation, unless the information is substantially easier to interpret in portrait orientation.
- Direct annotation of figures will be preferred to legends. All figures with more than one variable or item will contain either direct annotation or legends. All annotation will be unambiguously identifiable as such.
- Color will be used in figures only when needed to enhance clarity of communication. All color schemes will be evaluated for visual clarity for individuals with diminished color vision. All color encodings will be identified. Redundant encodings (such as the use of different plot symbols or line dash patterns) will be used in addition to color, so that all figures are interpretable after monochrome reproduction at 100 dots per inch. All dash patterns and line widths will be adequate to be distinguishable after monochrome reproduction at 100 dots per inch. Any distinction between plot symbols (circles, filled circles, diamonds, etc.) will remain clear after monochrome reproduction at 100 dots per inch.
- Fixed width sans serif fonts will be used for all labeling (Helvetica, Arial, or Futura).
- Boldface and italics will not be used unless substantial value is added.
- Decorative fonts and enhancements, including borders and shading, will not be used. Decorative presentation methods, such as ribbon graphs, will never be used.
- All information given in figures will also be presented in summary tables (perhaps only included in an Appendix or in supplementary materials).
- Only standard characters will be used in tables and data listings.
- All titles will be centered. The first title line will be the number of the table, figure, or listing. The second and possibly third lines will be the description of the table, figure, or data listing. The ICH numbering convention will be used for all.
- All footnotes will be left justified and at the page bottom. Footnotes will be used sparingly. Reference footnotes will be complete enough to locate any reference based on the information provided (Author, Journal, Pages, Date, or PubMed accession number).
- Missing values for numeric or character variables will be unambiguously identified as such using the special string `NA` (not available) in all settings; `NA` is the standard missing value code for our software. Each figure or table caption in which `NA` is used will indicate the meaning of `NA` in that figure or table. The abbreviation `NA` will never be used for any other purpose.
- All date values will be presented in the form DDmmmYYYY format (e.g. 01jan2008), using four digit years. June will be encoded as `jne` (otherwise `jan` and `jun` would differ by only a single character), and July as `jly` (so that the lowercase letter `l`, easily confused with the digit `1`, will not be adjacent to any numerals).
- All tables, figures, and data listings will have the name of the program and a date/time stamp on the bottom of the output.

# F Revisions

- October 29, 2020: Specified that time assessed as continuous variable in the primary analysis

# References

Chow SC, Shao J, Wang H. Sample Size Calculations in Clinical Research, Second Edition, Chapman & Hall/CRC, Boca Raton, 2007.

Deiner MS, Worden L, Rittel A, Ackley SF, Liu F, Blum L, Scott JC, Lietman TM, Porco TC. Short-term leprosy forecasting from an expert opinion survey. PLoS One. 2017 Aug 16;12(8):e0182245.

Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic evaluation in clinical trials. Oxford University Press, Oxford, 2007.

Lietman TM, Gebre T, Ayele B, Ray KJ, Maher MC, See CW, Emerson PM, Porco TC; TANA Study Group. The epidemiological dynamics of infectious trachoma may facilitate elimination. Epidemics. 2011;3(2):119-24.

Liu F, Porco TC, Amza A, Kadri B, Nassirou B, West SK, Bailey RL, Keenan JD, Lietman TM. Short-term forecasting of the prevalence of clinical trachoma: utility of including delayed recovery and tests for infection. Parasit Vectors. 2015;8:535.

Solomon AW, Holland MJ, Burton MJ, West SK, Alexander NDE, Aguirre A, Massae PA, Mkocha H, Muñoz B, Johnson GJ, Peeling RW, Bailey RL, Foster A, Mabey DCW. Strategies for control of trachoma: observational study with quantitative PCR. *Lancet* 262:198-204, 2003.

Salomon, JA, Vos, T, Hogan, DR, Gagnon, M, Naghavi, M, Mokdad, A, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. Lancet 2012; 380:2129-43.