STATISTICAL ANALYSIS PLAN for ACT4 Study

1. Data will be checked for completeness and cleaned
2. Descriptive Analysis-
   By arm, we will compare:
   - The number and size of sites, overall and by country in terms of:
     - Total number of new TB cases
     - number of microbiologically confirmed TB cases
     - number of clinical TB cases
   - The number and size of sites, overall and by country in terms of confirmed pulmonary TB cases overall and by age group
   - The number and size of sites, overall and by country in terms of number of contacts overall and by age category

Overall		Intervention	Control
	N TB index cases		
	N randomization sites	12	12
	Average number of TB index cases per randomization sites (SD)		
Country			
Canada	N TB index cases		
	N randomization sites	2	2
	Average number of TB index cases per randomization sites (SD)		
Benin	N TB index cases		
	N randomization sites	1	1
	Average number of TB index cases per randomization sites (SD)		
Ghana	N TB index cases		
	N randomization sites	1	1
	Average number of TB index cases per randomization sites (SD)		
Indonesia	N TB index cases		
	N randomization sites	4	4
	Average number of TB index cases per randomization sites (SD)		
Vietnam	N TB index cases		
	N randomization sites	4	4
	Average number of TB index cases per randomization sites (SD)		

3. The primary analysis will be an intention to treat analysis, using a Poisson regression approach. We will use a marginal Poisson regression model, estimated via GEE, and using an identity link [1] (See Appendix for details). We will use an exchangeable correlation structure at the level of the unit of randomization and use robust standard errors. Because the number of clusters is less than 40, we will use a correction for few clusters [2].

   The dependent variable will be the number of household contacts (HHC) who initiated treatment for LTBI per index TB patient (Y_i/TB_i). Two time points will be included – Phase 1: the 6-months before the program strengthening begins, and Phase 2: during the last 6-months, after program strengthening has been implemented. The model will include terms for the intervention, study phase, and the interaction between study phase and intervention. The interaction term, will be interpreted as the difference in the change from Phase 1 to Phase 2 in the number of HHC starting LTBI treatment per index TB patient between the intervention and control arms (i.e. a difference of differences). Hence, this interaction term is the primary focus of this analysis. Overdispersion will be investigated and accounted for if necessary [3].

4. In secondary analyses, we will consider:
   a. A Poisson regression model, with identity link, and separate fixed intercepts for each randomization unit, similar to equation 5 in the Appendix. This approach of Demidenko [4], is our first choice for secondary approaches, because this approach will allow us to describe the effect of the intervention in terms of a difference.
   b. A marginal Poisson regression model, with a log link, an exchangeable correlation structure at the level of the unit of randomization, with a correction for few clusters, and including log(TBi) as an offset. We will use robust standard errors. See equation 3 in the Appendix.
   c. A Poisson regression model with log link, and separate fixed intercepts for each randomization unit, including an offset.

5. In secondary analyses we will also do the following:
   - adjust the model for country income level (World Bank categories)
   - estimate the effect of the intervention separately in each country
   - evaluate the change in the outcome from baseline to follow up separately in experimental and control sites, using the same Poisson marginal model, with identity link, and correction for few clusters (see equation 6 in the Appendix).

5. In sensitivity analyses, we will consider as the outcome the proportion of identified, eligible household contacts who initiate LTBI treatment, using a binomial marginal model with an identity link (or logit link) estimated via GEE. In this case, the denominator is estimated and depends on the number of index cases with active TB, the average household size and the proportion of eligible contacts the program targeted. To account for the uncertainty in the

estimation of the denominator in sensitivity analyses, we will use a simulation based approach and generate the average household size and proportion eligible from country-specific distributions. By repeating this step, multiple times we can account for the uncertainty in the denominator.

REFERENCES:

[1] *Breslow, NE. Cohort analysis in epidemiology. From: Atkinson AC, Fienberg SE. A Celebration of Statistics, 1985, Springer New York, New York, NY, pages 109-143*).

[2] *Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials, JoAnna M Scott, Allan deCamp, Michal Juraska, , Michael P Fay, Peter B Gilbert, Stat Meth Med Res Volume: 26 issue: 2, page(s): 583-597)*.

[3] *Dean, Charmaine B., and Erin R. Lundy. "Overdispersion." Wiley StatsRef: Statistics Reference Online (2014): 1-9.)*

[4] *Demidenko International Statistical Review, v75 n1 (April 2007): 96-11*3

Appendix

Description of a Poisson model with an identity link:

As stated above, the dependent variable will be the number of household contacts who initiated treatment for latent TB per index case with active TB (Y_i/TB_i). Two time points will be included – Phase 1: the 6 month baseline period, before the program strengthening was started, and Phase 2: the last 6-months of the follow up period, post-program strengthening. The model will include terms for the arm (intervention vs. control), phase, and the interaction between phase and arm.

Usually, when fitting a Poisson regression model, the log link is used. In that case, we would fit:

$$\text{Log}(Y_i/TB_i)=\beta_0+ \beta_1*\text{Phase}_i+ \beta_2*\text{Intervention}_i+ \beta_3*\text{Phase}_i*\text{Intervention}_i \quad (1)$$

$$\text{Log}(Y_i) -\text{log}(TB_i)=\beta_0+ \beta_1*\text{Phase}_i+ \beta_2*\text{Intervention}_i+ \beta_3*\text{Phase}_i*\text{Intervention}_i \quad (2)$$

$$\text{Log}(Y_i)=\beta_0+ \beta_1*\text{Phase}_i+ \beta_2*\text{Intervention}_i+ \beta_3*\text{Phase}_i*\text{Intervention}_i+\text{log}(TB_i); \quad (3)$$

where Phase =0 if Phase 1, and 1 if Phase 2; intervention=0 for control health facilities and =1 for intervention sites; and, i denotes the randomization unit = 1, … 24.

Here, $\text{log}(TB_i)$ is known as the offset – it is a regression term that is forced in the model with a coefficient equal to 1 – notice that there is no β in front of it. It ensures that the interpretation of the model is in terms of the rate of contacts initiating treatment for LTBI per index TB case (Yi/TBi). When using a log link, the regression coefficients the log of the rate ratio.

However, we prefer to estimate a difference. In that case, instead of a log link, we use the identity link. The regression equation is thus:

$$Y_i/TB_i=\beta_0+ \beta_1*\text{Phase}_i+ \beta_2*\text{Intervention}_i+ \beta_3*\text{Phase}_i*\text{Intervention}_i \quad (4)$$

As described by Breslow, (*Breslow, NE. Cohort analysis in epidemiology. From: Atkinson AC, Fienberg SE. A Celebration of Statistics, 1985, Springer New York, New York, NY, pages 109-143*), we can multiply through by TB_i to obtain the model we fit:

$$Y_i =\beta_0*TB_i + \beta_1*\text{Phase}_i*TB_i + \beta_2*\text{Intervention}_i*TB_i + \beta_3*\text{Phase}_i*\text{Intervention}_i*TB_i \quad (5)$$

Notice that this model has no intercept, and that the terms included in the model are the number of index cases, the product of phase and number of index cases, the product of intervention and number of index cases, and the product of phase, intervention and number of index cases.

In either model (equation 3 or equation 5), the primary focus will be the interaction term β_3. In the model estimated using the log link (equation 3), $\exp(\beta_3)$ is interpreted as the rate ratio for intervention vs. control sites beyond any time trend. In the model estimated using the identity link (equation 5), β_3 is interpreted as the difference in the change from baseline to follow-up in the number of household contacts starting LTBI treatment per index case between the intervention and control arms (i.e. a difference of differences).

The log link is the canonical link and using it ensures that predictions made by the model will fall into the range of possibilities for a count (i.e. >0). When using the identity link, it is possible that the model predicts values <0. Thus, we will ensure that the predictions produced by the model are greater than 0

– if not or in the case of convergence problems, we will consider (i) a Poisson regression model with fixed intercepts for each site (*Demidenko International Statistical Review, v75 n1 (April 2007): 96-11*3); or (ii) using a log link as described above. The approach of Demidenko will allow us to describe the effect of the intervention in terms of a difference and so will be preferred.

In addition, we will evaluate the change in the outcome from baseline to follow up separately in experimental and control sites, using the same Poisson marginal model, with identity link, and correction for few clusters:

$$Y_i = \beta_0 * TB_i + \beta_1 * Phase_i * TB_i \quad (6)$$