Clean Data, Clean Hands:
Implementing Data Quality Assurance in the WASH Benefits Study

Eva Zenilman, Paul Byatta2, and Clair Null1
1Rollins School of Public Health-Center For Global Safe Water, 2Innovations for Poverty Action

The WASH Benefits Study

An estimated 2.2 million children under the age of 5 die each year from diarrheal disease, and this burden is believed to be preventable through WASH improvements (Ezzati et al., 2003)

Nutritionists have recently hypothesized that reducing a child’s fecal bacteria exposure during the first years of life may improve gut function and subsequent growth (Humphrey, 2009)

However, there is almost no evidence that allows direct comparison of the health benefits or cost-effectiveness of WASH improvements, nor on how the benefits of these interventions aggregate when provided in combination (WASH Benefits IRB 2012)

The WASH Benefits research study, funded by the Bill and Melinda Gates Foundation and implemented by Innovations for Poverty Action (IPA), focuses on water, sanitation, and hygiene (WASH) and nutrition in rural Western Kenya. Its goal is to generate rigorous evidence about the impacts of WASH and nutrition improvements on child health and development in the first two years of life, as this time period is critical for child growth, nutrient absorption, and cognitive development. WASH Benefits is unique in that it is the first WASH study to use randomized control trial methods at this scale to measure impacts of WASH and nutrition.

Data Sources

Village Census
Baseline Surveys & Sample Collection
Backcheck
Fidelity Assessment

1) In Bungoma and Kakagema Counties of Western Kenya, villages are censused - data is collected about the number of women pregnant, the number of children under the age of 5, and water source(s).

Village eligibility for the study is then determined.

2) IPA Enumerators administer a 2 hour survey to all pregnant women enrolled in the study. WASH Benefits aims to enroll 8,000-10,000 women. The survey contains over 3,000 variables. Samples of blood, stool, and drinking water are also collected for analysis.

The study enrolls women because the interventions are likely to have the biggest effect on the youngest children. The study will follow the newborn babies from birth to age 2 as they grow up in an environment that is hopefully cleaner thanks to the interventions.

3) At least 10% of study participants receive a second visit within a week after the baseline survey. This “backcheck” survey asks questions from the Baseline in which responses should be the same.

The backcheck survey is used for quality control of the enumerators.

4) Fidelity assessments are completed to assure that interventions (e.g. tippy taps, handwashing, Lipid-based Nutritional Supplementation) were delivered to the right people/households on the right timeline. The study also wants to ensure that study participants have sufficient interactions with health promoters to learn about the interventions.

Data Management

The project seeks to ensure that it collects and maintains data of the highest quality through its multiple data sources. All datasets, but especially those of this size, require constant oversight and clearly developed methodologies for ensuring the highest quality of data as its being collected. This is the goal of the Data Management Team.

Examples of My Role

1) How many respondents have given birth?
   Of the 471 Active Respondents in Block 1:
   -447 Babies Born
   -3 Miscarriages
   -22 Babies Deceased

2) What are the water sources we are working with?
   For the water arm, which is given a chlorine dispenser:
   -The majority of water sources are located in a private compound, as opposed to an open area, religious center, etc.
   -The types of water sources vary more: the four most common are a protected dug well, protected spring, unprotected well, and unprotected spring

3) Attrition: Is there any way to predict if respondents will drop out of the study?
   -39% of respondents (for block 1 and block 2) visited home relatives due to domestic dispute
   -100% of those respondents who visited home relatives due to domestic dispute are married and 71% of them own land
   -The majority of respondents who visited home relatives visited for a ceremony

Methods

1) Because survey questions can be misinterpreted by the enumerator and the respondent, finding how many babies have been born is not as straightforward as it seems. Respondents also could report false pregnancies. We created a do-file to that considers each month’s data.

2) I first created a CSPro data entry template and facilitated double data entry for the Chlorine Dispenser Eligibility and Information forms. I ran discrepancy checks and reconciled the data, as well as sorting through the duplicates. I then created a do-file to analyze the data.

3) One way to potentially answer this question is to look at the Marriage Module statistics, though I did not find a significant correlation between attrition and the information in the module when I ran a number of probit regressions. The Baseline Survey’s Marriage Module looks at a number of questions, such as if the respondent is married, who owns land and has paid dowry, and if the respondent has visited their home relatives and for how long and what reason.

Data quality is essential in order to make valid conclusions. In any complex project there are multiple threats to quality and even with the best designed systems, potential errors include reporting, transcription, and data entry errors. The QM/QA objective is to identify and correct these threats to data quality, and variable-by-variable review is one method to evaluate data consistency and quality across multiple data entry points.

In addition to the questions above, I helped in performing overarching tasks to ensure data quality. I created a codebook for the baseline survey and backcheck survey so anyone on the team can interpret the variables correctly. I created a new system for entering the data acquired in the Monthly Phone Calls made to each active household and created new data entry templates while reconciling the first four months of data. I also began to organize, enter, reconcile, and analyze data for the Intervention Delivery team, which collected data on Tippy Tap use and malfunctions. Variable-by-Variable review is also a consistent method to ensure data quality: I reviewed missing values, survey codes, numeric values (age, months pregnant), and string variables (names). [An example of this is shown below.] We also performed high frequency checks, such as enumerator productivity and logic patterns.

References:
WASH Benefits IRB approved 2012 and approved by Emory and KEMRI Kenya Medical Research Institute.

Acknowledgments:
I am grateful to Giti, Dr. Null, and Genesee Mason for giving me this research opportunity, and to Dr. Null, Paul, and Bonnie for welcoming me onto their team.