Distribution Agreement

Exploring changes in reported diarrhea among control group members in trials of water quality interventions

By

Grayson M. Privette
Master of Public Health

Global Environmental Health

Thomas F. Clasen, PhD
Committee Chair

Paige Tolbert, PhD
Committee Member

Exploring changes in reported diarrhea among control group members in trials of water quality interventions

By

Grayson M. Privette

B.A.,
The University of North Carolina at Chapel Hill
2010

Thesis Committee Chair: Thomas F. Clasen, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Environmental Health
2014

ABSTRACT

Exploring changes in reported diarrhea among control group members in trials of water quality interventions

By Grayson M. Privette

BACKGROUND Studies of the effect of water quality intervention on self-reported diarrheal disease have shown an apparent pattern of reduced levels of disease by members of the control group between baseline (or first surveillance point) and endline. If this change is statistically valid, it raises questions about the use of controls as a counterfactual to assess intervention effectiveness.

OBJECTIVE To determine whether and under what circumstances there are changes in reported diarrhea among control group members in trials of water quality interventions.

METHODS We compared levels of self-reported diarrhea at baseline (or first surveillance point) and endline from studies of water quality interventions. Studies were eligible for inclusion in the analysis if they were included in a 2014 systematic review of water quality interventions in low- and middle- income settings, the outcome was self-reported diarrhea and the required data was supplied by the study authors. Data was compiled on key study characteristics that could potentially explain changes in disease levels (study design, intervention type, length of follow up, etc.), presented descriptively, and analyzed using paired t tests.

RESULTS Of the 72 studies identified within the systematic review, 47 met the criteria for self-reported diarrhea as an outcome of interest. We were able to obtain data on beginning and end levels of diarrhea from 18 studies and included these in the analysis. Overall, there were no significant differences in the self-reported levels of diarrhea among control group members in water quality interventions from baseline (or first surveillance measure) to last surveillance measure. While differences were not statistically different, there was a consistent pattern of changes between beginning and end line outcomes, with 97% of the assessed relationships in this study showing a decline over time.

CONCLUSIONS Although no significant differences exist in reported diarrhea pre- and post-intervention, further research on this is warranted. A more comprehensive assessment of current water and sanitation literature may be able to elucidate the relationship suggested by these data of lower diarrheal disease reported among control group members in water and sanitation intervention trials.

Key words: water, sanitation, diarrhea, self-report, control group, paired t test

Exploring changes in reported diarrhea among control group members in trials of water quality interventions

By

Grayson M. Privette

B.A.,
The University of North Carolina, Chapel Hill
2010

Thesis Committee Chair: Thomas F. Clasen, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Environmental Health
2014

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Thomas F. Clasen for his time, effort and caring spirit throughout this compressed thesis time schedule. Also due an immense thanks is Mr. Paul Weiss, instructor in the Biostatistics Department at the Emory University Rollins School of Public Health—without his unceasing patience and philosophical discussion, this thesis project would not have come to fruition. I would like to cordially thank all of the researchers around the world that have assisted me with data collection and interpretation of results from their respective water quality interventions: Dr. Joe Brown, Dr. Sandy Cairncross, Dr. Ronan Conroy, Dr. Jan Graf, Dr. Josh Gruber, Dr. Jan Hattendorf, Dr. Mimi Jenkins, Dr. Louis Kirchhoff, Dr. Stephen Luby, Dr. Daniel Maeusezahl, Dr. Kevin McGuigan, Dr. Mark Sobsey, Dr. Christine Stauber, and Dr. Linda Venzcel for their communication, genuine interest and support of the project. I cannot forget the people who make these interventions necessary and worthwhile—those living in low- and middle-income countries serving as study subjects within water quality trials.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

BACKGROUND & INTRODUCTION

Causal Inference & Experimental Research

The key concept underpinning modern experimental research was an idea initially developed in the 17th and 18th centuries: to deliberately vary something so as to discover what happens to something else to learn about the effects of presumed causes (Shadish et al. 2002). To know if cause and effect are related, we borrow from the 19th century philosopher John Stuart Mill, whose analysis postulated (1) the cause precedes the effect, (2) the cause was related to the effect, and (3) there is no plausible alternative explanation for the effect other than the cause. These three tenets can be addressed through experimentation in which we (1) manipulate the tested cause and observe an outcome, (2) see whether variation in cause correlates to variation in effect, and (3) use multiple methods during experimentation to decrease the plausibility of other potential explanations for the effect, respectively (Shadish 2002).

Basic Components of Experimental Research

True experimental research has four basic components: random selection, random assignment, manipulation, and control (ORI 2014). The most important among these components are manipulation and control, as they are central to the idea of causal inference. Manipulation is a controlled change that is introduced by the researcher on the study population—the cause of the cause-and-effect relationship. Of utmost importance to this study, however is the control or comparison group, which helps to prevent outside factors from influencing the outcome, or effect of a study through an approximation of a valid counterfactual.

Having in place the appropriate components of a rigorous experimental study will allow for appropriate and accurate interpretation of intervention results. At the very basis of this concept, however, is that the comparison group provides an estimate of the background or expected exposure (case-control study) or incidence of disease (cohort study. Unfortunately, the ideal comparison group would consist of the exact same individuals as the treatment or experimental group, had they not been exposed. This model is called the counterfactual, or "contrary to fact" and it embodies what would have happened to the same population that received a treatment had they simultaneously not received the treatment.

Intervention Studies In Public Health Research

Intervention studies are among the most important tools public health researchers can use in the scientific quest to improve the human condition (NIH 2014). To some extent, all causal relationships are context-dependent, so generalization of an issue in experimental research is of particular concern especially in a field like public health where appropriate implementation and interpretation of research and experimental findings may safe millions of lives. In this vein, public health practitioners have utilized intervention studies to provide the scientific basis supporting major policy decisions and insights into focus areas such as healthy lifestyle promotion, prevention of disease transmission and best practices for disease treatment.

Put most simply, public health interventions such as the distribution of a water filter are experimental studies in which subjects undergo some kind of intervention—a new treatment program, drug, or surgery—in order to evaluate its impact (AFMC 2014). At its foundation, an intervention study involves formally stating the research hypothesis, objectives and outcomes of interest, random selection of study subjects from an eligible target population and enrollment at baseline, random allocation of the eligible subjects into two groups, delivery of the intervention(s) to the intervention group while withholding the intervention from the control group, and finally outcomes of interest that have been recorded over the study period are compared between the two groups (Talley et al. 1994; Fewtrell et al. 2005; Liu et al. 2012).

Intervention studies are, at their very heart, comparative evaluations of effect. Study subjects are selected from a particular population with a specific characteristic and are, immediately after baseline, separated into a group or groups that receive an intervention and a group that does not receive that intervention (Lesaffre et al. 2005). Such studies have been a standard instrument by which to improve scientific understanding in various disciplines within public health in particular, as they are the epidemiological studies that most closely resemble experiments conducted by scientists in the laboratory (WHO 1999).

Precise Estimates Require Precise Methodology

If this rigorous methodology is not followed or if the basic elements of the comparative evaluation of effect are not ideal, intervention studies can lead to spurious conclusions based on the presumption of scientific truth. Interventions with large measured effect sizes, for instance, have a much greater likelihood of leading to major changes in policy or increases in funding for that particular program. On the other hand, it is just as important to learn about interventions with smaller effect sizes so they can be identified and improved upon or abandoned. Hence, any over- or under-estimation of an effect size must be minimized in order to provide the best evidence for policy and programming in the future.

Additional methodological importance is placed on calculations for study sample size for adequate statistical power in epidemiological studies and with intervention studies in particular. Such calculations involve assumptions about the underlying prevalence of disease in the study population and are essential in gaining a sufficient study population to adequately address your research hypothesis. It is crucial, therefore, that the assumptions being made and followed in experimental research be accurate and appropriate for the quality of methodology that is employed. The implications of such research are simply too important to rely on assumptions, such as a control group approximating a valid counterfactual when they do experience intervention effects, that may well not be accurate in reality.

Control Group Assumptions

At the core of a valid impact evaluation using an experimental design are several assumptions about the control group that provide greater assurance that measured effects are due to the presence or absence of the intervention itself and not some outside factor. One assumption of control groups is baseline equivalence with the intervention group. Equivalence is the idea that study groups are not statistically different in any major characteristic of interest except for the provision or absence of the intervention—something that can only be accomplished through group randomization (Kapthuk 2001).

Another assumption of the effect estimations determined through group comparisons is that, without the intervention, the control group is not affected by the intervention in any way, an approximation of the counterfactual ideal. Valid comparison groups must involve tests for statistical equivalence between the treatment and control groups that show they approximate one another before the intervention begins.

In intervention studies, the development of a counterfactual is accomplished by using a control group created through random assignment from our eligible study population so they should not be statistically different from our intervention group. Although there is debate as to the feasibility of such estimations, as in complex situations it can be impossible to determine an accurate estimate of a counterfactual measure of the outcome, it is nonetheless an ideal guide to follow for the control group. In sum, the two central foci for experimental intervention design are the creation of a high-quality but realistically imperfect source of counterfactual inference (control group) and the understanding of how this referent group differs from the treatment group (Shadish 2002).

In some public health interventions, however, all basic methodology can be followed meticulously, but the control group may still move away from its ideal representation of the counterfactual. For example, there may be a spillover of beneficial intervention effects if a particular intervention works to decrease the exposures. Say the two experimental groups are made up of subjects living in the same village, and the intervention group is receiving a latrine as part of this intervention trying to improve access to sanitation. If an intervention is effective in reduction of diarrhea, say because the intervention member now uses the latrine versus opting for open defecation, the reduced disease levels in the intervention population may be such that the environmental dispersion of enteric pathogens is decreased. This herd effect is not something that reflects poorly on the validity of the effect, but is merely a by-product of reduced exposures outside the realm of the intervention.

Misclassification of a control group or individual, on the other hand, is a reflection on the validity of the study itself. In intervention studies, misclassification of the study participants involves the erroneous classification of study arm members to the wrong arm. This phenomenon may happen, for example, where community level interventions in a village improve that community's pumped water supply. The control member is somehow able to gain access to this water either by visiting the pump or a neighbor's house that has access to the pump, and is thereby granting them access to the intervention treatment (de Heer et al. 2011). Where members of the control group are somehow receiving the intervention through their daily activities or changes in behavior that would otherwise classify them as an intervention group member, their allocation to the control group results in misclassification (Jepsen et al. 2004). Depending on the direction of the drift, misclassification may work to amplify or attenuate the magnitude or level of impact of an intervention treatment over the course of follow-up.

Enrollment and Measurement at Baseline

Another critical component of study design that may impact the control group is the enrollment and baseline measurement period. In this time frame, much before the initiation of the intervention, study investigators are in the process of identifying eligible and willing study participants through the use of informed consent. This process provides insight for each study member into what the research question and general experience will be like as a control or intervention member. With this preview comes the potential expectation for compensation or some useful item or treatment out of the process. When this does not happen or when it does, the feelings generated either by an absence of incentives or the promise of incentives can both profoundly affect subjective recall measures such as self-reported diarrhea (Clasen, TF, personal communication).

The establishment of a baseline rate of diarrhea and assessing experimental group characteristic equivalence is crucial to all epidemiologic research, as it provides the reference basis for both internal and external validity. Internal validity is the ability to say whether or not observed co-variation in a condition of interest between the study arms should be interpreted as a causal relationship, which is only possible if good baseline measurements are taken to characterize the study population itself (Lesaffre et al. 2005, Campbell and Stanley 1966). It is also important to determine the degree of this equivalence at baseline, ideally to ensure both study groups are comparable but also to adjust results estimates accordingly. External validity is the ability to say whether the causal relationships found in a particular intervention can be generalized to different measures, persons, settings and times, and it also relies upon measurement at baseline to characterize potential differences between the study sample and a wider population (Steckler et al. 2008).

Baseline measurements in trials serve to verify randomization success, adjust the final analysis for imbalances between study arms and increase the precision of the treatment effect by inclusion of the baseline measure as a covariate in an adjusted analysis (Schmidt et al. 2011). Trial reports usually include the baseline clinical characteristics of interest from study groups as these measures form the basis upon which to assess for the external validity by a comparison with other settings and other populations (Rothwell 2005). It is important to note, however, that all studies do not undertake or report baseline data, instead relying upon random assignment to study groups to ensure equivalence between the groups (internal validity) and random selection to ensure external validity or comparability with the general population (Rothwell 2006).

Diarrheal Disease and Applications to WASH Intervention Studies

An important application of intervention trials and measures of effect can be seen in water, sanitation and hygiene (WASH) interventions to reduce the global burden of diarrheal disease. Diarrheal disease is among the leading causes of mortality among children under five worldwide, responsible for 760,000 deaths in children under 5 and 1.4 million total deaths every year (Lim et al. 2012; WHO 2013). The burden of morbidity attributable to diarrhea in terms of DALYS (disability-adjusted life years lost) is also high (Liu et al. 2012; WHO/UNICEF 2013; CDC 2013). Children under three years old in developing countries experience an average three episodes of diarrhea annually and form a significant portion of the nearly 1.7 billion annual cases of diarrhea globally (WHO 2013).

There is evidence that a significant proportion of morbidity and mortality associated with diarrheal disease can be prevented through safe water quality and adequate sanitation and hygiene. (Fewtrell et al. 2005; WHO 2008; Cairncross et al. 2010) Despite considerable worldwide efforts and attention, however, the obstacles remain great; an estimated 786 million individuals still utilize unsafe water quality and 2.5 billion lack access to improved sanitation (WHO/UNICEF 2013).

Among the causes of diarrheal disease include inadequate water supply, low water quality, poor access to sanitation, and lack of hygiene education (Fewtrell et al. 2005). Since the mid-1980s, a variety of water and sanitation (WASH) intervention trials have been published in an attempt to determine the most effective and efficient means by which to meet these critical needs through delivery of water and sanitation services, infrastructure and/or education to the populations most in need (Fewtrell et al. 2005). Although many of these interventions have been reported to be successful in reducing diarrhea (Fewtrell et al. 2005, Cairncross et al. 2010), most of the underlying studies rely on a subjective or 'soft' measure: self-report of diarrhea (Cairncross et al. 2010). In numerous trials of water quality interventions with self-reported diarrhea as the outcome, control groups are reporting lower levels of diarrhea following the baseline. On the other hand, objective or "hard" measures, such as microbiological indicators of fecal contamination or residual free chlorine in treated water can serve as a benchmark measure for intervention effectiveness and can play a compelling role in improving water quality (Clasen et al. 2007, Cairncross et al. 2010).

Defining Diarrhea

It is crucial for WASH intervention studies with diarrhea self-report as an outcome to provide a clear definition of diarrhea to its study population as well as one that is consistent with international guidelines (Talley et al. 1994 & OTHERS). Diarrhea is defined as the passage of three or more loose or liquid stools per day, or more frequent passage than is normal for the individual (WHO 2013). The frequent passing of formed stools and of so-called "pasty" stools by breastfed babies should be differentiated from diarrhea as it is necessary to accurately define this symptom as an infection in the intestinal tract that may have multiple causal determinants.

Metrics and Biases of Diarrheal Disease Research

Once a concise case definition of diarrhea has been established, adequately described and is well understood by study participants, data collection can begin and the intervention may be evaluated on its effect on each study group. The best methods to ascertain the most accurate level of illness in a study population, however, are subject to a series of biases and can be difficult given the complex nature of description necessary for data collection of diarrhea (Schmidt et al. 2011). There are several important positive and negative attributes of a diarrheal self-report that are essential to understand in order to take the collected data and make an informed determination of an intervention's effect. Diarrhea self report is a known, albeit subjective, measure that has been subject to a range of validation studies (Ramakrishnan et al. 1998; Arnold et al. 2013; Feikin et al. 2010). It is also one that does not require extensive laboratory input and allows a description, with some degree of accuracy, of the true levels of illness in a community (Schmidt et al. 2011). Currently there is no objective measure that has been validated by epidemiologists thus far that is practical for large-scale implementation (Null et al. 2009).

Potential and inherent biases with self-reported diarrhea are many, including the over-report of diarrhea to gain entry into a study, observer effect, recall bias and respondent fatigue among others (Ramakrishnan et al. 1998; Arnold et al. 2013; Feikin et al. 2010; Schmidt et al. 2011). If there are perceived benefits to being a part of an intervention, study participants may over-report their level of diarrhea in an attempt to improve their odds of inclusion. Reactivity or "Hawthorne effect" refers to a phenomenon where study subjects alter their behavior as a result of being part of a study. This effect may alter the behavior of both intervention and control groups and may be related to the duration and frequency of follow-up visits as prescribed per the study design.

Recall bias in diarrhea self-report is another major potential obstacle to ascertaining true disease levels, as inaccuracies in diarrhea recall are directly related to the length of the recall period and inversely related to the severity of diarrhea (Alam et al. 1989; Ramakrishnan et al. 1998). The measurement of diarrhea is most commonly utilized is self-reported incidence of episodes over a specified period of time, usually a 3- or 7-day recall period (Schmidt et al. 2011). There is some guidance about recall length for optimal disease reporting accuracy for mothers, caregivers and self-report. There is an underestimation of severe cases of diarrhea with weekly recall survey data collection as well as, to a lesser extent, less severe cases and there is considerable underreporting of diarrhea morbidity when recall periods exceed three days. Research has recommended collecting this information on a shorter 3- to up to 7- day basis to yield more accurate data, convenience (Ramakrishnan et al. 1998; Arnold et al. 2013; Feikin et al. 2010).

Respondent fatigue is another obstacle to accurate collection of diarrhea self-report data (Null et al. 2009) and is otherwise known as a “bugger-off” effect—a phenomenon in which participants, especially control group members, alter their reporting to what they believe will satisfy intervention data collectors as they become increasingly disillusioned with a long length of follow-up and/or high frequency of follow-up visits (Clasen TF 2013). In fact it has been shown that a high frequency data collection appears to lead to unreliable measures of child diarrhea prevalence (Null et al. 2009, Zwane et al. 2011).

The assessment of baseline measures as a basis for the evaluation for internal and external validity, the standardized measurement and reporting of diarrheal disease and an understanding of inherent biases all play a role in the overall evaluation of WASH interventions and their effectiveness.

Current Examples from the Literature

In some trials of water quality interventions with self-reported diarrhea as the outcome, however, control groups are reporting dramatically lower levels of diarrhea following the baseline. The described phenomenon can be better explained with the assistance of a few examples from recent water quality literature. Looking at the Boisson et al. 2013 trial in Orissa, India (figure 1 below), we see a precipitous decrease in control group reported levels of disease followed by a longer, more gradual decrease in reported diarrhea over the course of this double-blinded intervention's follow-up period. What is also interesting to note is how closely the control mirrors the reported rates from the intervention group during the entire yearlong study. It does not seem, from this visual analysis, as if the control is approximating the counterfactual model as desired.

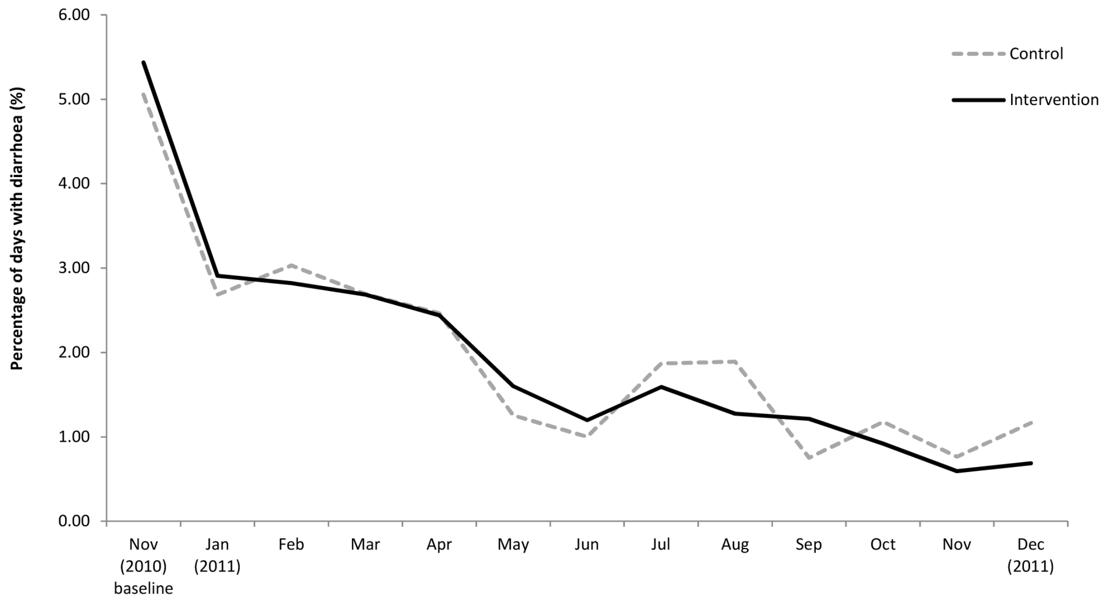


Figure 1. From: Boisson et al. 2013. “Effect of household-based water quality chlorination on diarrhoea among children under five in Orissa, India: a double-blind randomised placebo-controlled trial.” PLoS Med. 2013 August; 10(8): e1001497. “Figure 2. Prevalence of diarrhoea among children <5 y over time.”

Another example, figure 2, comes from the Jain et al. 2010 triple-blinded, placebo controlled study of sodium dichloroisocyanurate tablets in Ghana. Again, we see a close mirroring of the intervention group and a gradual decrease in reported diarrhea rates among control group members over time.

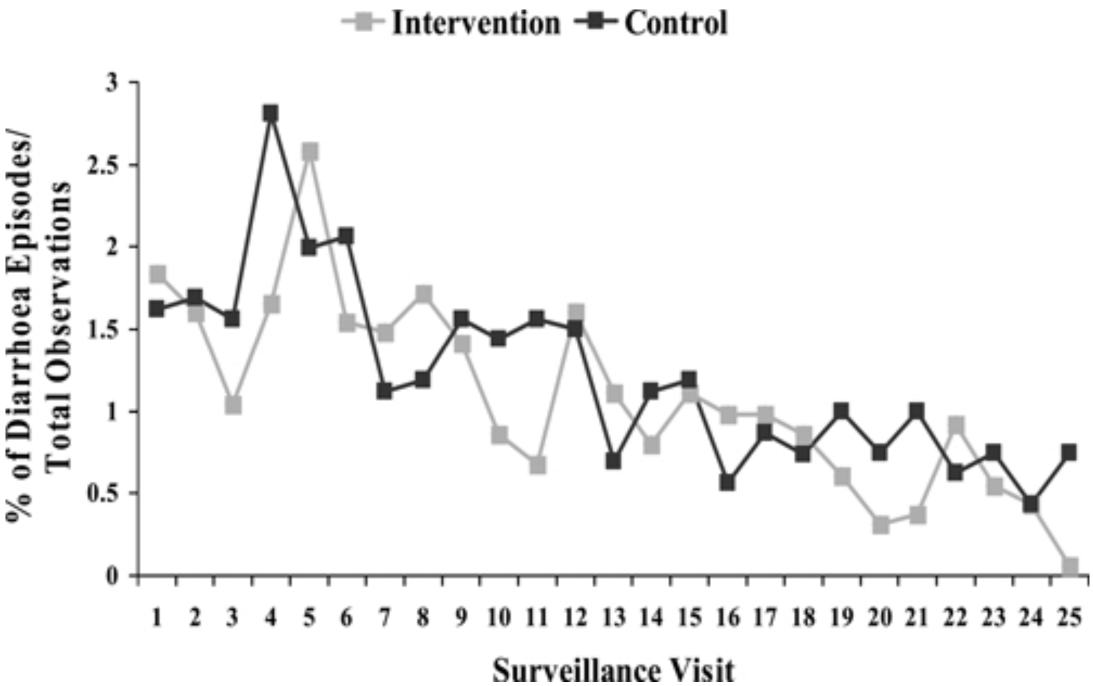


Figure 2. From: Jain et al. 2010. Sodium dichloroisocyanurate tablets for routine treatment of household water quality in periurban Ghana: a randomized controlled trial. Am J Trop Med Hyg. 2010 January; 82(1): 16–22. "Percent of diarrhea episodes per total number of observations in intervention and control groups, by surveillance visit (*N* = 3240)."

Figures 3 and 4 below present additional examples of this apparent phenomena. Here, an additional period of measurement before the intervention was initiated—a multi-week and multi-month surveillance period to capture baseline diarrheal rates among the Quick et al. 1998 and Stauber et al. 2012 interventions, respectively.

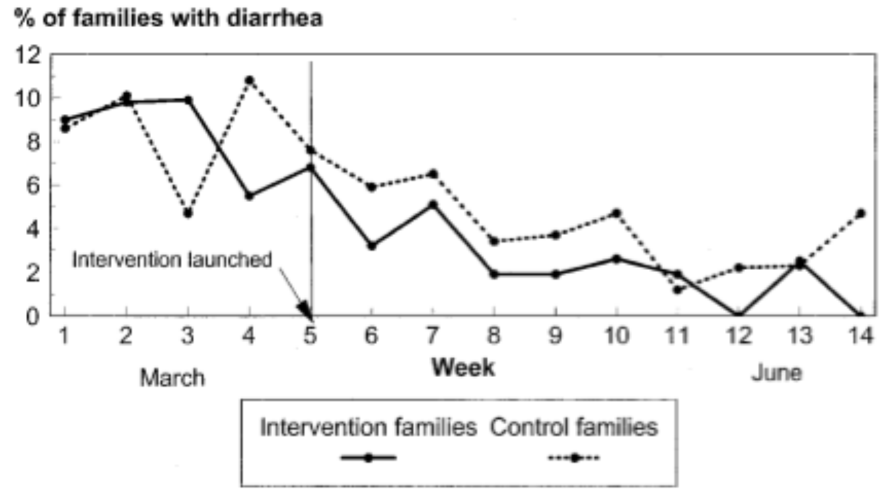


FIGURE 2. Percent of households in intervention and control groups with at least one diarrhea episode, by week of study, Kitwe, Zambia, March-June, 1998.

Figure 3. From: Quick et al. 2002. "Diarrhea Prevention Through Household-Level Water Disinfection and Safe Storage in Zambia. *Am. J. Trop. Med. Hyg.,* 66(5), 2002, pp. 584–589

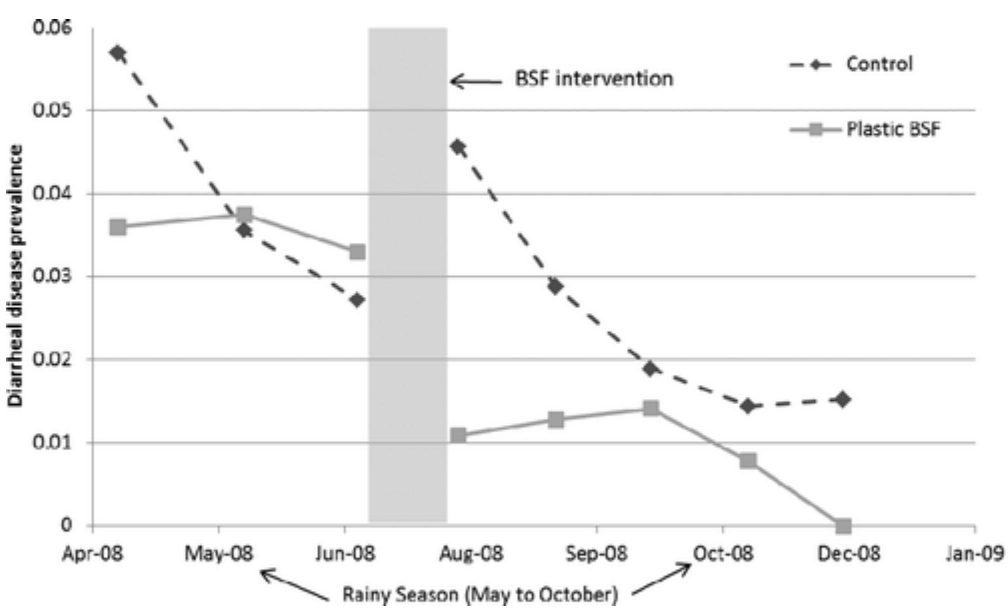


Figure 4. From: C. E. Stauber et al. 2012 "Cluster Randomized Controlled Trial of the Plastic BioSand Water Filter in Cambodia." *Environ. Sci. Technol.*, 2012, 46 (2), pp. 722–728. "Figure 2. Monthly point prevalence of diarrheal disease for plastic BioSand Water filter (BSF) and control groups over the period April–December 2008. The shaded area indicates the timing of plastic BSF installation."

Investigators involved in the research have noted how this pattern seems to persist in many intervention studies of WASH interventions where the outcome is reported diarrhea; they describe it as the "hockey stick" (TF Clasen, personal communication). To date, however, whether this pattern is actually common across these studies, and if so the factors that may explain it, has never been analyzed. The underlying assumption is that the level of disease in the controls will not be affected by the trial because they approximate the valid counterfactual. Effect estimates of these interventions in comparison to the control are used to direct evidence based decision-making in global health policy and programming. If this phenomenon is common, it may be the result of important yet often overlooked biases in water quality intervention trials. A descriptive analysis of this phenomenon is an important first step in terms of learning about the magnitude and true direction of this effect in hopes of recognizing and being able to account for it in the future.

RESEARCH QUESTION

In trials of water quality interventions that use self-reported diarrhea as the primary outcome, is there evidence that the control group is reporting less disease than would be expected from a valid counterfactual?

OBJECTIVE

To determine whether and under what circumstances there are changes in reported diarrhea among control group members in trials of water quality interventions.

SPECIFIC AIMS

- To identify water and sanitation intervention studies that record self-report of diarrhea as an outcome.
- To collect data on the level of self-reported diarrhea at baseline (or in the absence thereof, first surveillance point) from such intervention studies.
- To compare control group baseline (or first surveillance point) disease frequency with endline to assess intervention effects.
- To briefly consider characteristics of intervention studies that are associated with any unexplained changes in the level of disease reported by the control group.

METHODS

Initial Selection of Studies

Data for this study was drawn from trials of water quality interventions that used reported diarrhea as an outcome. In an effort to include as many relevant studies as possible in this analysis, we selected studies from the most recent systematic review of water and sanitation interventions to prevent diarrhea (Wolf et al. 2014). The review, which was led by the World Health Organization, included 72 experimental studies of water and sanitation interventions conducted from 1985 to 2013. That review was based on a comprehensive search of five databases, including Cochrane Library, MEDLINE and

PubMed, Global Health, Embase and BIOSIS in May 2013, and a specific search term strategy was followed (Wolf et al. 2014). In accordance to selection criteria for the review, eligible study designs included randomized controlled trials (including cluster randomized controlled trials), quasi-randomized and non-randomized controlled trials when baseline information on self-reported diarrhea was available pre- and post-intervention, case-control and cohort studies, observational studies using specific matching methods, and combined implementations (Wolf et al. 2014).

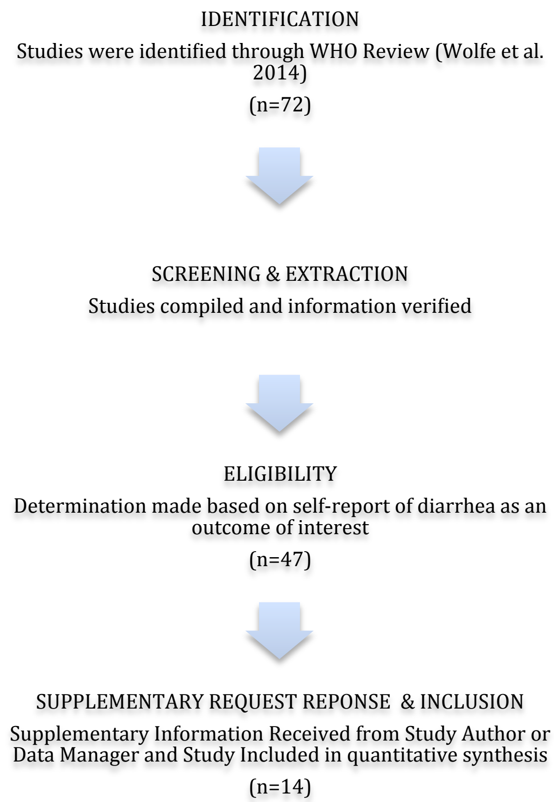
The review excluded studies that targeted healthcare facilities, schools or the work place, used special population groups (i.e. HIV-positive persons), had self selection into the study population or used interventions that were not acceptable to the population (implementation rates lower than 20%). (Wolf et al. 2014) Additional characteristics of the literature include the requirement to be published in a peer-reviewed journal or whose methodological quality was assessed according to transparent criteria in a previously published systematic review.

The author reviewed the titles and abstracts of all 72 studies included in Wolf et al 2014 to determine whether self-reported diarrhea incidence or longitudinal prevalence was as a primary or secondary outcome. Key data and study characteristics, including any measures of effect recorded on diarrheal disease, were then extracted directly from the published studies and organized by the author into a database.

Contacting researchers to obtain missing data

Few studies included all the required data, especially raw point data for multiple time points within the intervention baseline and follow-up period. A thorough examination of eligible literature revealed that characteristics on reported rates of diarrhea for the experimental groups, other than summary measures in water and sanitation intervention trials was very difficult to ascertain. Furthermore, rates on the control group are seldom reported with the same frequency as the intervention group if these rates are recorded at all. It was determined that a comprehensive data collection form was necessary to create in order to solicit the necessary raw supplementary data from researchers while simultaneously being the least intrusive on their time.

Contact information was obtained for lead and secondary authors of each study. Supplementary data requests were then performed through email for each of the studies through a data extraction form that was created specifically for this study (APPENDIX). The next step involved sending supplemental data request forms attached to emails to the most recent emails available found for each study author through a web search. The initial round of messages did not solicit enough of a response for adequate data collection, so another round of messages were sent to many of the same study authors to request these data. Studies were excluded if selected study authors failed to respond and/or were unable to provide or obtain the necessary data from their intervention.



Basic flow of the selection process for water quality interventions

Data extraction

Critical data for each supplementary raw data request included several important variables. The author was asked to review and approve of the basic study data already included on the request form, including authorship, year of publication, year of intervention, country of implementation, and a brief description of the intervention type.

The portion of the form to be filled in by the researcher requested a brief description of the measure of disease frequency, and a series of questions to ascertain whether a baseline and/or first surveillance measure was taken for both the intervention and control groups within the intervention. Specifically, data on baseline (or first surveillance period) diarrhea incidence or longitudinal prevalence for control and intervention groups was requested for each study of interest in addition to endline diarrhea rates for control and intervention. There was a question regarding whether group equivalence between baseline and control was established at baseline, if assessed. There was also an additional notes section attached to the database form to provide room if the study author wished to add additional study details or data qualifications and clarifications. Although diarrhea was classified as the passage of 3 or more loose or liquid stools per day, consistent with the current WHO/UNICEF standard definition (WHO 2005), alternative definitions of diarrhea were permitted if the specific intervention definition was assessed for validity. These data formed the basis for exploring the fundamental underlying assumption of the comparison group neutral counterfactual, that the level of disease in the control group was not affected by the trial.

Statistical Analysis

Data Preparation

Once a completed supplementary data request form was received, the data was aggregated within a master dataset including over 40 characteristic data variables for each intervention. Difference estimates were performed for the reported diarrhea rates among the control groups by quantifying the percent change in whichever measure was recorded in the intervention for at least two of the three time points of interest for this study—baseline or first disease surveillance or follow-up measure, and final follow-up measure. The rates reported at these time periods were paired with the paired observation from the same control group at a later time period. This difference was divided by the first measure to yield a percent difference in reported diarrhea.

An important caveat to this standardization of effect through a calculation of percentage change was a consideration for the extensive heterogeneity of the analyzed water quality interventions. Although standardization of intervention effect was attempted on a trial-by-trial basis, this percent change calculation could not account for all potential confounders as discussed later in this analysis. The specific measures of disease frequency ranged from longitudinal or period prevalence measures to incidence rates and rates of disease per a certain amount of person-time.

The author recognizes many of the inherent difficulties in the aggregation and analysis of these data together and would mention that this is being performed solely for the purposes of this initial, descriptive study of control group effects. After the database was complete, these raw data were entered into SAS for initial cleaning and initial, descriptive univariate analyses.

Data Analysis: Paired T-Testing

When data is collected at two points in time—usually baseline and follow-up—a paired t-test may be conducted to determine the significance of the differences in the sampled data over this time period to provide the basis for causal inference of the observed effect (Hsu et al. 2008). The paired t test pairs two observations from the same individual or group so each serves as its own control and so avoiding variation among groups. After the data for this study was obtained from responding study authors, it was sorted and compiled in a database and subsequently categorized. The three primary comparisons of interest for significant difference analysis were from baseline to endline (last surveillance point, first surveillance point to endline and finally first measure (a combined baseline and first surveillance point) to endline.

The equation for the paired t test is as follows:

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

In the formula above, the matched pairs must be between pairs of observations matched into meaningful groups, in this study the same group is being measured over two time points and is paired with itself. The average (X_D) and standard deviation (s_D) of the differences measured between these two groups are used in the equation where the constant μ_0 is non-zero since we're testing to see if the mean of the difference is significantly different from μ_0, with n-1 degrees of freedom. Hence, a normal distribution is a required assumption of the data for this formula to be both appropriate and accurate in its evaluation of the differences observed between matched observations (Zimmerman 1998, Pappas 2004).

The first step in analysis involved calculating univariate statistics of interest to determine the assumption of normality of these data. The specific statistics most crucial for this appraisal of the data are the simple histogram plot, mean and median relationship, skewness statistic and kurtosis statistic. If the data was deemed to follow a normal distribution for this small sample dataset, paired t testing would commence. If the skewness and kurtosis parameters fell outside of the acceptable range of -1 to 1 for the approximate normal distribution or if the distribution was observed to have significant outliers and/or a heavy tailed distribution, there was a significant possibility that the data are non-normally distributed. If this is the case, there are two options available: a data transformation was performed first to potentially correct for this non-normality or a different nonparametric testing procedure was utilized. (Northwestern 2014, Zimmerman 1998).

Intervention Characteristic Stratification

The impetus behind of intervention sub-grouping for stratified analysis is to test for multiple variables that may be of importance for the outcome of interest. In water quality interventions, there are many variables that are potentially at play in a particular exposure-disease or disease-exposure relationship, such as the relationships between water quality and diarrheal disease. For the researcher, it is useful to separate groups by the outcome or grouping of a particular variable to see if there are any significant changes in the outcomes for each group when analyzed separately. Put more simply, it is the examining of the exposure-disease association within different categories of a third factor.

Stratified analyses can help to disentangle issues around confounding and effect modification by looking at the effects of two different variables on disease at the same time (McNamee 2003 and Gregg 2008). Deciding on variables to stratify often involves an understanding previous literature that suggests a relationship or potential confounding with a particular variable or characteristic. It also may involve looking more closely at a variable that in previous studies has a suggested or known clinical importance.

Studies were grouped into sub-categories according to a number of study design characteristics deemed to be potentially important by the author and the research team to reported rates of diarrhea in the control group. As there was no literature specific to control group effects of intervention trials upon which to build strata, the decision to stratify on a particular characteristic hinged upon the suggested meaningfulness of the data along with its availability from the literature data extraction or supplementary requests. Characteristics included the level of the intervention treatment (point-of-use or community) protective effect, blinded status, diarrhea recall period, study design and establishment of group statistical equivalence at baseline. Characteristics of intervention type included type of point of use water treatments such as chlorination, filtration and solar disinfection and community-level treatments such as piped water infrastructure were analyzed separately whenever possible to ascertain intervention-level differences in modeling.

Table 1. Relevant Drinking-Water Intervention Characteristics for Paired T-Testing Analyses

Characteristic	**Investigated Stratum**
Treatment Type	Point-of-use: Chlorination Filtration Solar disinfection Community-level: Piped supply
Study Design	Cross-sectional survey, household-level randomized controlled trial, individual-level randomized controlled trial, non-randomized controlled trials, prospective cohort)
Length of Follow-up	Number of Months
Frequency of Data Collection Visits	Visits per Month of Follow-up
Cumulative Data Collection Visits	Total number
Diarrhea Recall Period	3-day, 7-day, or Other
Study Setting	Urban or Rural
Effectiveness of the Intervention	Proportional to estimate of effect size

A final major cluster of analysis was performed on continuous characteristics of the follow-up period for each intervention. Some researchers have suggested the possibility of a “bugger-off” effect of interventions that have longer durations, greater frequency of follow up or a greater number of total visits during the study period. The hypothesis is used to try to explain reported decreases in the level of disease among control groups over the follow-up period even if they may not be receiving any measurable positive effect of an intervention. Intervention quality was not considered as a criterion for stratification or characteristic of interest for this study. Although quality assessment measures were made by Wolf et al. 2014, time constraints did not allow for an in-depth analysis of the quality assessment measures used. To have greater control over what variables had an effect on analysis, it was therefore deemed appropriate to exclude the quality assessment measures from any stratification analyses.

Unless otherwise noted, all statistical analyses were performed by use of SAS software (version 9.3, SAS Institute Inc., Cary, NC.)

RESULTS

Sources of Data

From the 72 water quality and sanitation intervention trials included in a recent World Health Organization meta-analysis of water and sanitation interventions in low- and middle income settings, 47 were deemed to contain the appropriate information about self-reported diarrhea as an outcome of interest (Wolf et al. 2014). The other 25 intervention trials were deemed inappropriate for this analysis due to the lack of self-reported diarrhea as an outcome of interest. After screening the full text of these 47 studies to confirm that they met the eligibility criteria, we attempted to contact primary or secondary authors from each study for supplemental information. An example template for supplementary information is included in the Appendix.

After several weeks of follow up with study authors, we obtained responses for 33 intervention trials. Authors of 14 trials did not respond to the inquiry. Of the authors who responded for the 33 intervention trial data requests, authors of 15 studies reported that the required data was not available due to changing locations, busy schedules, an unsuccessful archive search, previous data destruction or that the study was not appropriate for this analysis template. In the end, 11 different authors provided requested data from 18 separate intervention trials in order to be included in the analysis. Figure 5 provides a review of the study flow displaying the study inclusion and data analysis inclusion processes.

Given the small relative number of studies in this review, it was more practical to perform analyses combining effects of multiple point-of-use treatment types as well as study designs. There were three blinded study designs that both met the inclusion criteria and from which supplementary data was received for this study (Kirchoff et al. 1985; Jain et al. 2010; Boisson et al 2013). These studies were not separated from other study designs due to their small relative number.

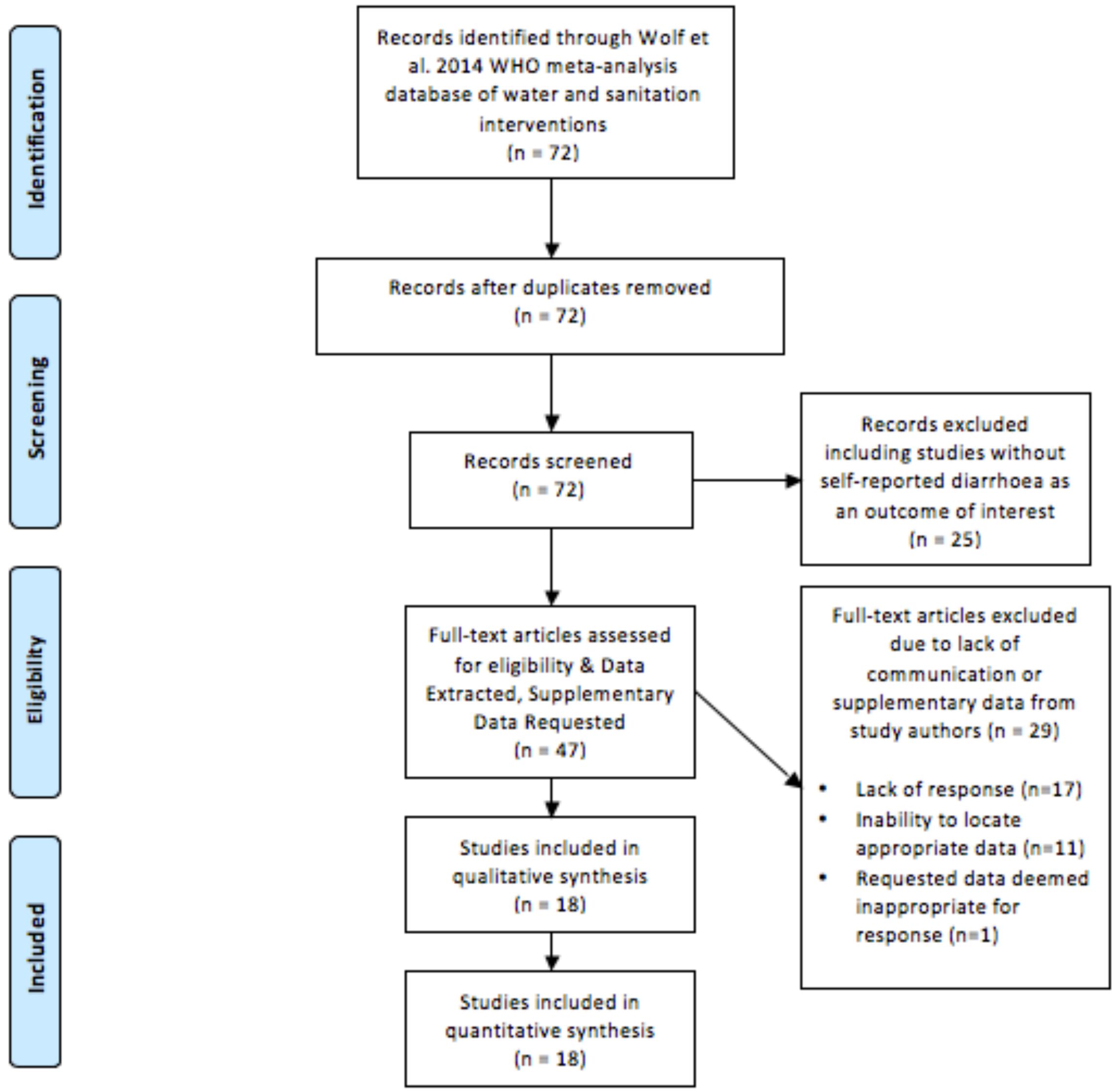


Figure 5. Study flow diagram

Description of Included Studies

Types of Interventions

Among the 18 studies included for qualitative and quantitative analysis, 2 were community-level piped water interventions and 16 were various point-of-use (POU) or household-level interventions of varying type. Thus, none of the studies were of sanitation interventions even though the Wolf review included 11 studies of sanitation and 61 studies of water quality as these interventions either did not contain data on self-reported diarrhea or requests for supplementary data were unfulfilled. The interventions described in the included studies were published between the years 1998 and 2013. All were published, following the inclusion criteria for the Wolf et al. 2014 meta-analysis (Table 2).

Table 2. Selected Descriptive Characteristics

Author	Title	Publication	Year of Intervention	Country	Type of Intervention
Boisson S, Stevenson M, Shapiro L, Kumar V, Singh LP, Ward D, Clasen T.	Effect of household-based water quality chlorination on diarrhoea among children under five in Orissa, India: a double blind randomised placebo-controlled trial.	PLoS Med. 2013; 10(8): e1001497.	2010-2011	India	POU*, chlorination
Brown J, Sobsey MD, Loomis D.	Local water quality filters reduce diarrheal disease in Cambodia: a randomized, controlled trial of the ceramic water purifier	Am J Trop Med Hyg 2008; 79: 394–400.	2007	Cambodia	POU, ceramic filter
Brown, J., V. T. Hien, L. McMahan, M. W. Jenkins, L. Thie, K. Liang, E. Printy and M. D. Sobsey	Relative benefits of on-plot water supply over other 'improved' sources in rural Vietnam	Trop Med Int Health 2013; 18(1): 65-74	2012	Vietnam	Piped water to premises
Clasen T, Garcia Parra G, Boisson S, Collin S.	Household-based ceramic water filters for the prevention of diarrhea: a randomized, controlled trial of a pilot program in Colombia	Am J Trop Med Hyg 2005; 73: 790–5.	2004-2005	Colombia	POU, ceramic filter
Clasen TF, Brown J, Collin S, Suntura O, Cairncross S.	Reducing diarrhea through the use of household-based ceramic water filters: a randomized, controlled trial in rural Bolivia	Am J Trop Med Hyg 2004; 70: 651–7.	2003	Bolivia	POU, ceramic filter

Clasen TF, Brown J, Collin SM.	Preventing diarrhoea with household ceramic water filters: assessment of a pilot project in Bolivia	Int J Environ Health Res 2006; 16: 231–9.	2004	Bolivia	POU, ceramic filter
Graf, J., S. Z. Togouet, N. Kemka, D. Niyitegeka, R. Meierhofer and J. G. Pieboji	Health gains from solar water disinfection (SODIS): Evaluation of a water quality intervention in Yaounde, Cameroon	J Water & Health 2010; 8(4): 779-796.	2007	Cameroon	POU, solar treatment, hygiene education
Jain, S., O. K. Sahanoon, E. Blanton, A. Schmitz, K. A. Wannemuehler, R. M. Hoekstra and R. E. Quick	Sodium dichloroisocyanurate tablets for routine treatment of household water quality in periurban Ghana: a randomized controlled trial.	Am J Trop Med Hyg 2010; 82(1): 16-22.	2006	Ghana	POU, chlorination
Luby SP, Agboatwalla M, Hoekstra RM, Rahbar MH, Billhimer W, Keswick BH.	Delayed effectiveness of home-based interventions in reducing childhood diarrhea, Karachi, Pakistan.	Am J Trop Med Hyg 2004; 71: 420–7	2000-2001	Pakistan	POU, chlorination, safe storage, hygiene
Luby SP, Agboatwalla M, Painter J, et al.	Combining water quality treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial.	Trop Med Int Health 2006; 11: 479–89.	2003	Pakistan	POU, chlorination, flocculation, safe storage
Maeusezahl D., Christen A., Pacheco GD., Tellez FA., Iriarte M., et al.	Solar water quality disinfection (SODIS) to reduce childhood diarrhoea in rural Bolivia: a cluster-randomized, controlled trial.	PLoS Med 2009; 6: e1000125.	2008	Bolivia	POU, solar treatment, hygiene education
Majuru, B., M. M. Mokoena, P. Jagals and P. R. Hunter	Health impact of small-community water supply reliability	Int J Hyg & Env Health 2010. 214(2): 162-166.	2007-2008	South Africa	Community piped water supply
Quick, R. E., A. Kimura, A. Thevos, M. Tembo, I. Shamputa, L. Hutwagner and E. Mintz	Diarrhea prevention through household-level water disinfection and safe storage in Zambia.	Am J Trop Med Hyg 2002; 66(5): 584-589.	1998	Zambia	POU, chlorination, safe storage, hygiene education

Reller, M. E., C. E. Mendoza, M. B. Lopez, M. Alvarez, R. M. Hoekstra, C. A. Olson, K. G. Baier, B. H. Keswick and S. P. Luby	A randomized controlled trial of household-based flocculant-disinfectant water quality treatment for diarrhea prevention in rural Guatemala	Am J Trop Med Hyg 2003; 69(4): 411-419.	2001-2002	Guatemala	POU, chlorination, flocculation
Sobsey, M. D., T. Handzel and L. Venczel	Chlorination and safe storage of household water quality in developing countries to reduce waterborne disease.	Water Sci & Tech 2003; 47(3): 221-228.	2002	Bangladesh & Bolivia	POU, chlorination, safe storage
Stauber, C. E., E. R. Printy, F. A. McCarty, K. R. Liang and M. D. Sobsey	Cluster randomized controlled trial of the plastic BioSand Water filter in Cambodia	Env Sci & Tech 2012; 46(2): 722-728.	2008	Cambodia	POU, biosand filter
Stauber, C. E., G. M. Ortiz, D. P. Loomis and M. D. Sobsey	A randomized controlled trial of the concrete biosand filter and its impact on diarrheal disease in Bonao, Dominican Republic.	Am J Trop Med Hyg 2009; 80(2): 286-293.	2005-2006	Dominican Republic	POU, biosand filter, safe storage
Tiwari, S. S. K., W. P. Schmidt, J. Darby, Z. G. Kariuki and M. W. Jenkins	Intermittent slow sand filtration for preventing diarrhoea among children in Kenyan households using unimproved water sources: Randomized controlled trial	Trop Med Int Health 2009; 14(11): 1374-1382.	2007	Kenya	POU, biosand filter

*POU=point-of-use

Of the 18 studies included in the analysis, there were 7 (38.9%) chlorination; 7 (38.9%) filtration; 2 (11.1%) flocculation; 5 (27.8%) safe storage of water quality; 2 (11.1%) solar disinfection (SODIS); 4 (22.2%) hygiene; and 8 (44.4%) combination intervention types, respectively (Table 1). The studies evaluated interventions on three separate continents: Asia (4), Southeast Asia (3), Africa (5), South America and the Caribbean (7). They were conducted in a total of 14 different countries; Bangladesh, Bolivia (4), Cambodia (2), Cameroon, Colombia, Dominican Republic, Ghana, Guatemala, India, Kenya, Pakistan

(2), South Africa, Vietnam, and Zambia (Table 1). Of note, one study (Sobsey et al. 2003 in Bolivia and Bangladesh) reported on water quality interventions in two countries.

Study Designs

Of the 18 studies included in the analysis, there were 14 (77.8%) randomized, controlled trials (RCTs); 2 (11.1%) non-randomized controlled trials, 1 (5.6%) longitudinal prospective cohort study and 1 (5.6%) cross sectional survey design. Of the interventions included in this study, 2 were blinded interventions (11.1%) and 16 were non-blinded interventions (88.9%) (Table 3).

Table 3. Additional Selected Descriptive Characteristics

Author & Year	Design	Measure of Self-Reported Diarrhea
Boisson S et al. 2013	RCT	Longitudinal Prevalence*
Brown J et al. 2008	RCT	Longitudinal Prevalence°
Brown, J et al. 2013	Longitudinal Prospective Cohort Study	Longitudinal Prevalence°
Clasen TF et al. 2005	RCT	Period Prevalence
Clasen TF et al. 2004	RCT	Prevalence
Clasen TF et al. 2006	RCT	Prevalence
Graf, J et al. 2010	Two cross-sectional surveys	Prevalence
Jain, S et al. 2010	RCT	Incidence
Luby S et al. 2004	Non-RCT	Incidence
Luby S et al. 2006	RCT	Longitudinal Prevalence
Maeusezahl D. et al. 2009	RCT	Incidence
Majuru, B et al. 2010	Non-RCT	Incidence
Quick, R et al. 2002	RCT	Prevalence
Reller, M et al. 2003	RCT	Incidence
Sobsey, M et al. 2003	RCT	Incidence
Stauber, C et al. 2012	RCT	Incidence
Stauber, C et al. 2009	RCT	Incidence
Tiwari, S. S et al. 2009	RCT	Longitudinal Prevalence

RCT=Randomized, Controlled Study

* Defined as (#days with diarrhea / total number of days of observation).

¢ 7-day recall diarrheal disease (yes/no)

Outcome Measure

Incidence was the outcome measure for self-reported diarrhea in 7 (38.9%) of the 18 studies considered in this analysis (Table 3). Prevalence, most often longitudinal prevalence, was the measure of self-reported diarrhea in 11 (61.1%) of the 18 analyzed interventions.

Intervention Effectiveness

Of the included studies, 3 (16.7%) reported interventions that were ineffective (non-protective) while 15 (83.3%) were considered effective interventions with a statistically significant protective effect.

Control and Intervention Baseline Equivalence

As stated by the responding study authors and upon review of each intervention (n=18), the control and intervention groups were considered equivalent at baseline in 12 (66.7%) water quality interventions. These two study groups were not statistically equivalent at baseline in 3 (16.7%) of the 18 studies. There were also 3 (16.7%) interventions for which no data was received on this statistic.

Measured Time Points

For 13 (72.2%) interventions, the baseline measure of control group self-reported diarrhea was reported. For 8 (44.4%) interventions, the first measure reported was the first surveillance measure of the control group. There were also 3 (16.7%) interventions for which both the baseline and first surveillance measure was obtained. All 18 studies reported an endline or last surveillance point for the control group, as requested by the author of this study.

Supplemental data gathered from study authors from each of the included water quality interventions are listed in Table 4. These data are the basis for analyzed effects in this study in that they show the magnitude and type of change in self-reported rates of diarrhea within the control groups over the intervention. The combined, complete list of first reported measures (either baseline or first surveillance point) were utilized for the creation of the variable "first measure".

Table 4. Summary of Supplementary Data*

Author & Year	Controls at Baseline	Controls at First Surveillance Measure	Controls at Last Surveillance Measure	Percent Change in Controls (Base or First to Last)
Boisson S et al. 2013	0.051	.	0.012	- 76.951
Brown J et al. 2008	0.180	.	0.120	- 33.333
Brown, J et al. 2013	.	0.048	0.048	0
Clasen TF et al. 2005	0.240	.	0.096	- 60.069
Clasen TF et al. 2004	0.221	.	0.230	4.054
Clasen TF et al. 2006	0.067	.	0.047	- 29.906
Graf, J et al. 2010	0.343	.	0.318	- 7.288
Jain, S et al. 2010	.	0.016	0.008	- 53.703
Luby S et al. 2004	.	0.456	1.870	310.087
Luby S et al. 2006	.	0.028	0.035	25.266
Maeusezahl D. et al. 2009	6.860	.	3.870	- 43.586
Majuru, B et al. 2010	0.800	.	0.620	- 22.500
Quick, R et al. 2002	0.067	.	0.048	- 28.358
Reller, M et al. 2003	.	4.5	2.980	- 33.778
Sobsey, M et al. 2003	4.000	.	2.500	- 37.500
Stauber, C et al. 2012	0.035	0.039	0.015	- 57.265
Stauber, C et al. 2009	0.025	0.061	0.029	14.000
Tiwari, S. S et al. 2009	0.022	0.146	0.088	293.586

*Note: "•" = missing data

Subject Characteristics

Participants within these studies were of all ages from low- and middle-income settings and reported diarrheal disease morbidity without concern for etiology and case confirmation. Although some studies focused on the clinically important age grouping of children less than five, these interventions were reported along with interventions that randomized entire villages for the purposes of this analysis. The WHO review focused its eligible studies on intervention trials in low- and middle-income countries.

Characteristics of study follow up

Table 4 shows selected characteristics of the follow up strategy of each of the 18 studies included in the analysis. As previously reported, the arithmetic mean follow-up duration in this group of interventions was measured to be 32.11 weeks. The longest follow-up period grouping among the group of analyzed studies ranged from 52 to 56 weeks or slightly longer than a one-year period, which was found in three studies (Majuru et al. 2010; Reller et al. 2003; Boisson et al. 2013). Alternatively, the shortest follow-up period grouping ranged from 12-14 weeks (Jain et al. 2010, Quick et al. 2002).

The arithmetic mean number of follow-up visits to study households or individuals in this group of analyzed interventions was 25.35. The greatest number of overall visits characterized the Luby et al. 2006 intervention in Pakistan, where study workers visited participants 74 times over its duration. The fewest number of overall follow-up visits was 4 over the duration of the study; a characteristic shared by three studies—Clasen et al. 2004, Clasen et al. 2005 and Brown et al. 2013.

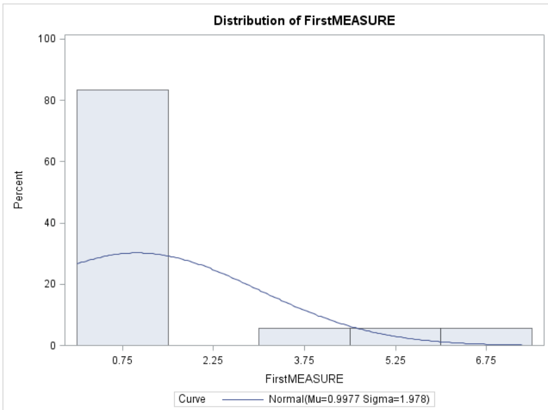
The arithmetic mean frequency of follow-up visits, calculated in visits per month, was 3.41 for this group of analyzed water quality interventions. Follow-up visit frequencies ranged from 0.72/month (Clasen et al. 2004, Clasen et al. 2005) to 8.7/month (Luby et al. 2006, Jain et al. 2010). These measures are summarized in Table 5.

Table 5. Continuous Variable Characteristics, Water quality Interventions

Author	Weeks of Follow-up	Number of Visits	Frequency of Visits (Monthly)
Boisson S et al. 2013	52.176	12	1
Brown J et al. 2008	22	11	2.174
Brown, J et al. 2013	17.392	4	1
Clasen TF et al. 2005	24	4	0.724666667
Clasen TF et al. 2004	24	4	0.724666667
Clasen TF et al. 2006	21.74	7	1.4
Graf, J et al. 2010	43.48	.	.
Jain, S et al. 2010	12	24	8.696
Luby S et al. 2004	23.914	24	4.348
Luby S et al. 2006	37	74	8.696
Maeusezahl D. et al. 2009	52	55	4.583333333
Majuru, B et al. 2010	56	56	4.348
Quick, R et al. 2002	14	15	4.658571429
Reller, M et al. 2003	52.176	52	4.348
Sobsey, M et al. 2003	30.436	30	4.348
Stauber, C et al. 2012	26.088	12	2
Stauber, C et al. 2009	43.48	40	4
Tiwari, S. S et al. 2009	26.088	7	1

Tests for Normality

We assessed the assumption of normally distributed data using a variety of univariate statistics from the primary outcomes of interest (baseline, first surveillance, first —combined baseline and first— and last surveillance time point measures), including skewness, kurtosis and the comparability of the median and mean. It was determined, after this review, that the data were non-normally distributed. A histogram overlaid with the normal curve and probability plot for the time point variable, first measure, is included in Figure 6 for review.



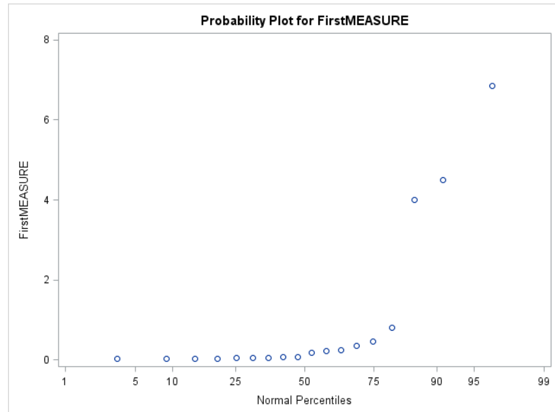


Figure 6. A histogram of the sample distribution overlaid with normal distribution and a probability plot of a selected variable of interest, first measure.

The univariate data of interest in the assessment of normality are the mean/median relationship, skewness and kurtosis of the sample distribution. For the variable first measure*,* although the mean (0.998) and median (0.124) are fairly close to one another, the skewness (2.24) and kurtosis (4.17) are both outside the -1 to 1 criterion range for the normal distribution. The data are right skewed and highly asymmetric.

Log-Transformation of the Data

Since the data were determined to be non-normally distributed, a transformation, or a single function applied to the data, was applied to correct for the non-normality to allow for the appropriate application of the paired t testing in subsequent analyses. A logarithm or log-transformation was selected because it can help to reduce right skewness and is especially appropriate in this analysis because it can "pull in" values greater than 1 relative to values less than 1, which is useful in correcting skewness to the right (Northwestern, 2014).

A paired sign test could have been performed with these data, as the direction of the differences in time point measures for diarrhea are known and it does not assume a symmetric population distribution for the paired differences. The situations in which this test is the most powerful is if the distribution was extremely heavy tailed, a characteristic that does not describe the time point measures in this analysis. Hence, a transformation of the data to correct for non-normality was selected over the paired sign testing.

After log-transformation, descriptive univariate statistics for the time measures of interest much more closely resembled the normal distribution. The mean (-1.77) and the median (-2.21) remain approximately similar to one another while the skewness (0.74) and kurtosis (-0.54) are, in effect, "pulled in" to within the -1 to 1 criterion range characterizing an approximately normal distribution. The results of this log-transformation are graphically appraised with the selected variable, first measure, in Figure 7. Here we can see the effect of the transformation in correcting for the previous non-normality of the original data. The effect of log-transformation was that the data now approximate the normal distribution, thereby allowing for the appropriate use and interpretation of paired t testing for these reported rates of diarrhea among control group members across the duration of water quality interventions.

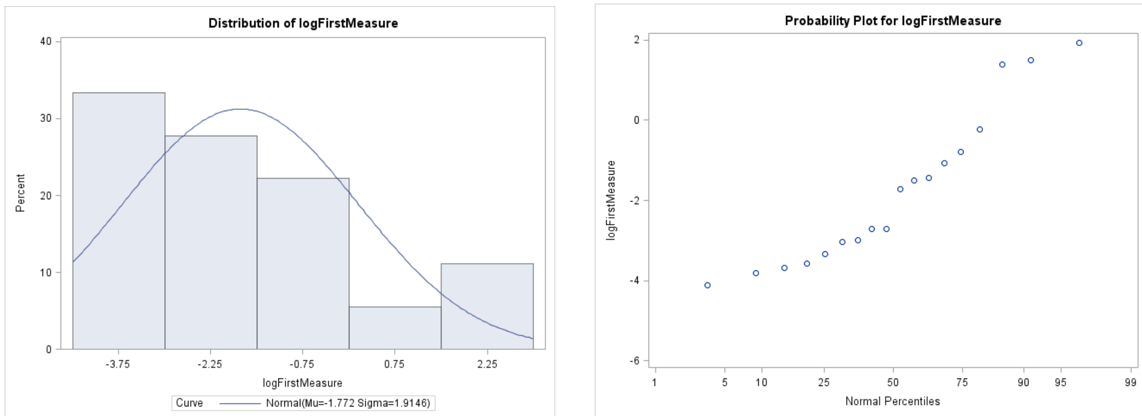


Figure 7. The sampling distribution overlaid with normal curve and probability plot for the log-transformation of the selected variable, first measure.

A visual method of analyzing characteristics of the sample distribution for the paired t testing in each of the three time points within water quality interventions is the Quantile-Quantile (or Q-Q) plot of difference. The Q-Q plot assesses the normality assumption for the differences in rates of self-reported diarrhea among the different time points. Figure 8 shows how the points are distributed about a line that represents the normal distribution in the relationship of baseline to endline reported diarrheal rates. Although our original points were highly non-normal, the log transformation of this distribution is shown to approximate a normal distribution as the points themselves are roughly distributed about the line at near-45-degree angle. An approximate appraisal can also be made of the log-transformed distribution of the sample distribution of paired observations as it relates to the first surveillance measure to endline (figure 9) and the first measure (combined baseline and first surveillance measure) to the endline (figure 10). All three of these figures work to tell us that the log-transformation of this distribution has worked to correct the previous non-normality and allow for paired t testing assumptions.

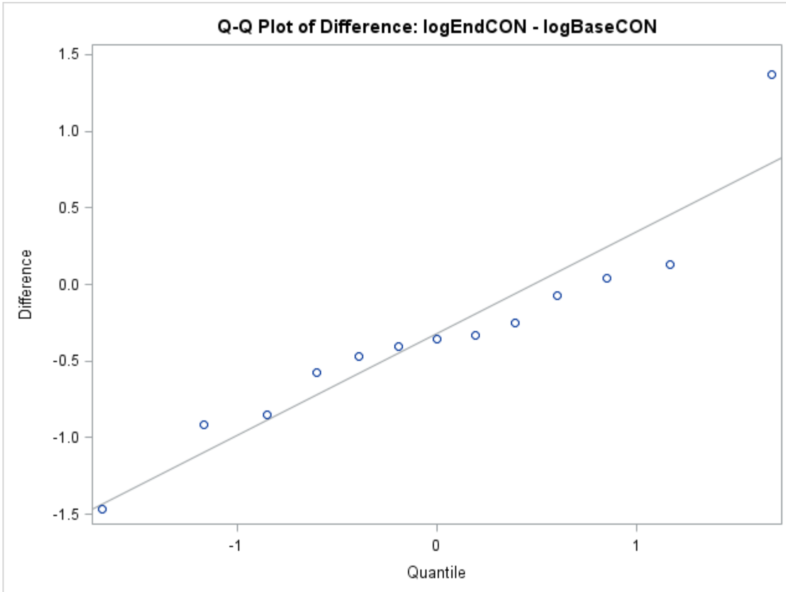


Figure 8. Paired T-Test Results Comparing Log-Transformed Baseline Measure to Log-Transformed Final Surveillance Measure. The Q-Q plot assesses the normality assumption for the differences in rates of self-reported diarrhea among the different time points. The figure shows how the points are distributed about a line that represents the normal distribution. As you can see, these points are highly non-normal.

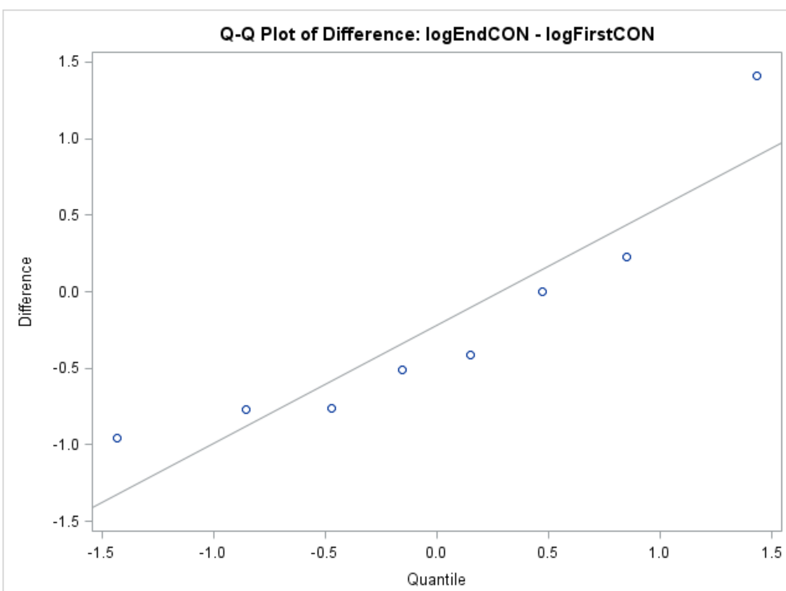


Figure 9. Paired T-Test Results Comparing Log-Transformation of First Surveillance Measure to Log-Transformation of Final Surveillance Measure. The Q-Q plot assesses the normality assumption for the differences in rates of self-reported diarrhea among the different time points. The figure shows how the points are distributed about a line that represents the normal distribution. As you can see, these points are highly non-normal.

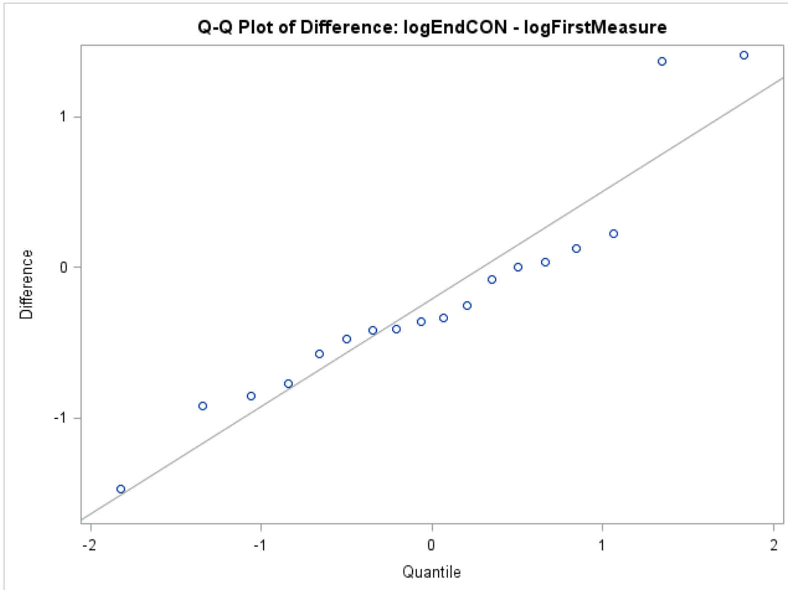


Figure 10. Paired T-Test Results Comparing Combined Baseline and First Surveillance Measure to Final Surveillance Measure. The Q-Q plot assesses the normality assumption for the differences in rates of self-reported diarrhea among the different time points. The figure shows how the points are distributed about a line that represents the normal distribution. As you can see, these points are highly non-normal.

Reported Differences in Control Group Diarrhea

There were three separate measurement point analyses to explore the significance of changes experienced by control groups over the course of water quality interventions. The three time point relationships that were explored included baseline to final surveillance measure, first to last surveillance measure and a combined first measure (baseline and first measure) to final surveillance point. Insignificant effects on self-reported diarrheal rates within control groups characterized all three relationships over the analyzed interventions. Each of these three paired t tests, however, reported a negative effect measure (Table 6).

Table 6. Results of Log-Transformed ($\log_{10}$) Paired T-Testing for Control Groups with Different Time Points (n=18)

Control Group Time Point Comparison	No. Interventions	Mean Log-Difference (95% CI)	P-Value$^{\infty}$
I. Baseline to Last Surveillance ($\log_{10}$)	13	-0.320 (-0.722, 0.082)	0.108
II. First Surveillance to Last Surveillance ($\log_{10}$)*	8	-0.221 (-0.866, 0.423)	0.444
III. Baseline or First Surveillance to Last Surveillance ($\log_{10}$)	18	-0.206 (-0.561, 0.149)	0.238

*Supplementary data for first surveillance point was received in addition to baseline measures for some studies.

Of the 18 interventions analyzed by this study (Figure 11), there were 5 (27.8%) that recorded an increase in reported diarrhea among the control group over the course of the intervention. Twelve (66.7%) of the 18 water quality intervention trials recorded a decrease in reported diarrheal disease levels among the control group. There was one study that recorded zero change in the control group over the course of the intervention (Brown et al. 2013). Although under the threshold for significance within the realm of acceptable statistical findings, the observation that studies that show the endline at a lower level than a first measure suggests a pattern that across this small sample of water quality intervention trials.

Mean changes in reported rates of diarrhea among control group members in the 18 analyzed interventions are not significantly different from zero, so we are unable to rule out the null hypothesis. The mixed-measure data on changes in diarrhea reported by control groups among these water quality interventions from first measure (baseline or first surveillance point) to last measure are displayed in Figure 11. The data in this spaghetti plot reflect the incidence and prevalence measurements as noted previously.

The general trend is clear: many control groups in water quality interventions report lower levels of disease from pre- to post-intervention. This idea is further exhibited with figure 12, which shows the magnitude and direction of percentage changes in reported diarrhea by control group members in these water quality interventions.

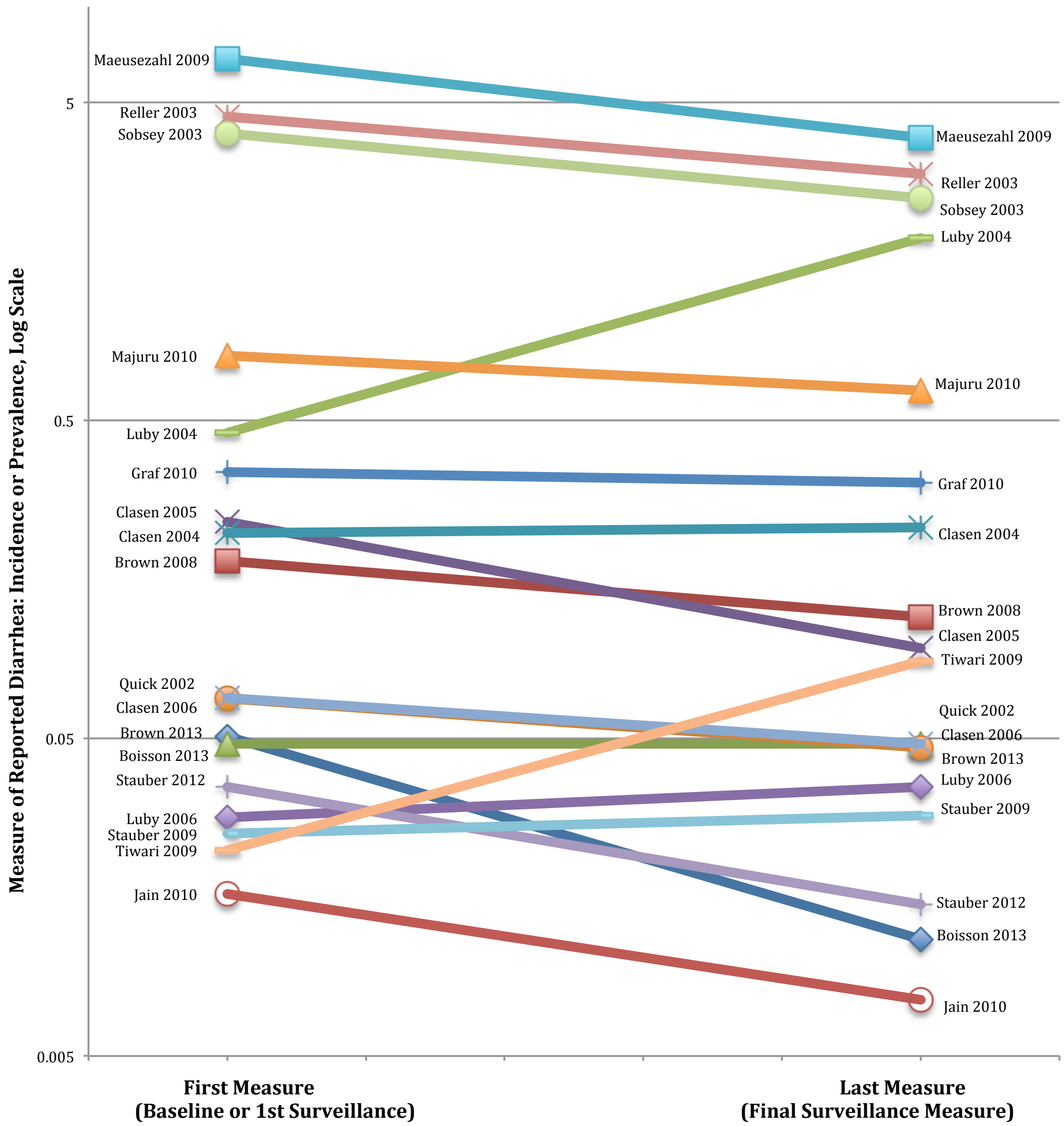


Figure 11. Mean Changes in Reported Diarrhea Rates Among Control Group Members in Water quality Intervention Studies; first measure (Baseline or First Surveillance) to Last Measure (Final Surveillance). Changes reported over the intervention period for water quality interventions (n=18); non-transformed mixed methods (incidence and prevalence) rate calculation; reported on a log-scale.

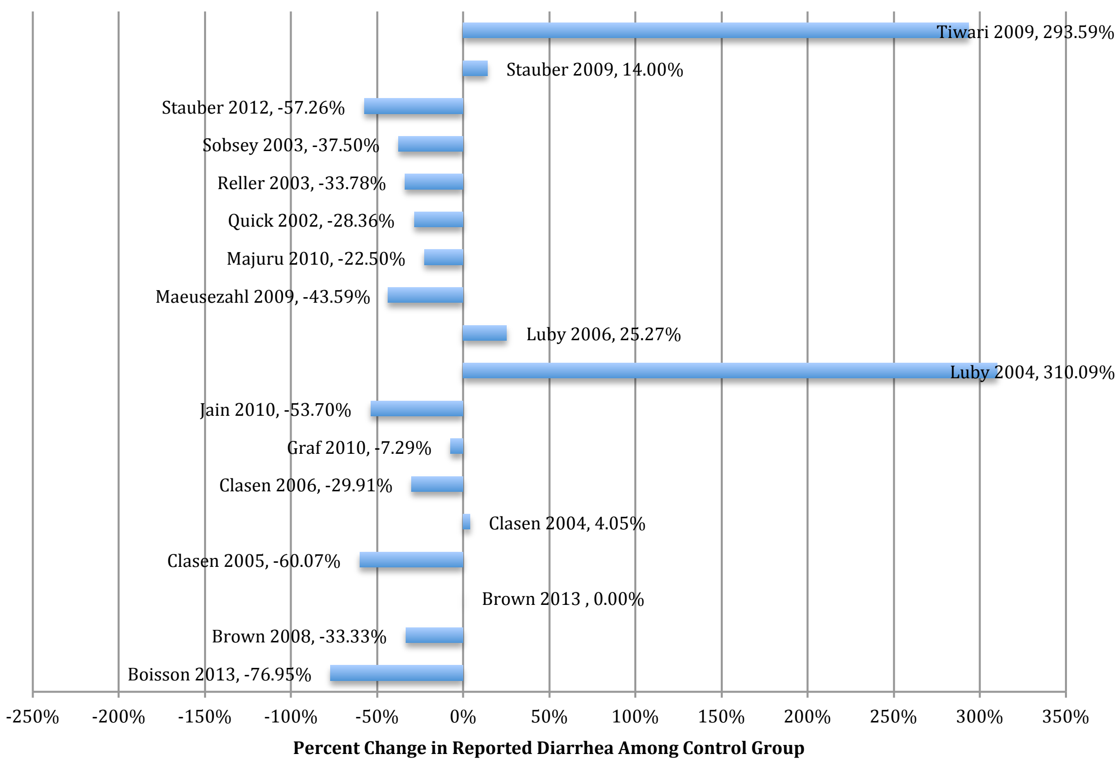


Figure 12. Percentage change in reported diarrhea among control groups of water quality interventions (n=18), from first measure (baseline or first surveillance point) to last measure (final surveillance point)

On average, the self-reported rates of diarrhea of control subjects in these studies did not have a significant measure of effect before versus after the intervention. The paired t-tests showed no significant mean log-differences were reported by the control group in each of the studies from baseline to last surveillance point, (-0.320 (-0.722, 0.082); p=0.108), first surveillance point to last surveillance point (endline) (-0.221 (-0.866, 0.423); p=0.444), or from baseline to first surveillance point (-0.206 (-0.561, 0.149); p=0.238), respectively. In summary, there was not a statistically significant difference in the rates of self-reported diarrhea after the intervention when compared to reported rates before the intervention.

Stratified Analyses

All stratified t-test results were observed to be insignificant with respect to different strata of the characteristic or variable of interest. No significant mean log-differences were discovered in any of the different levels of intervention characteristics of interest including study type or design, length of follow-up, total number or cumulative data collection visits. The first sets of stratified analyses were performed to test aspects of study design and methodology (Table 6). The second grouping of analyses focused on continuous variable characteristics of the intervention trial follow-up period. The final cluster of analyses elucidated effects of intervention type—the direct methods and processes for improving water quality—within this relatively small sample population of studies.

Of the 34 total stratified analyses performed by this study, 33 (97.1%) reported a negative (yet insignificant) effect on rates of reported diarrhea among control groups. The only effect measure that was calculated to be positive was observed in the non-RCT designs, with a mean log-percentage effect of 0.270 (-0.952, 1.493) and a p-value of 0.53 (Table 6). A summary of the paired t-testing effect measures, confidence intervals and associated p-values for all general and stratified analyses are displayed in tables 6-8.

Aspects of study design and methodology did not seem to have a significant effect on reported rates of diarrhea among control group members in water quality intervention trials (Table 6). Such important characteristics as blinding status, utilization or non-utilization of the WHO definition of diarrhea, differential recall periods and intervention effectiveness were among intervention methodological and impact characteristics that were observed to have insignificant effects on control group self-reported rates of diarrhea. The closest to borderline significance achieved by stratified paired t-testing analyses was an intervention having an RCT study design (p=0.076) and having no significant protective effect (p=0.075).

Table 7. Results of Stratified Paired T-Testing By Intervention Characteristic

Characteristic	No. Studies	Mean Log-Difference (95% CI)	P-Value
Study Design			
RCT	(14)	-0.342 (-0.726, 0.042)	0.076
Non-RCT	(4)	0.270 (-0.952, 1.493)	0.533
Blinding Status			
Blinded	(2)	-1.119 (-5.550, 3.312)	0.192
Non-Blinded	(16)	-0.092 (-0.444, 0.261)	0.587
WHO-defined Diarrhea			
Defined	(16)	-0.201 (-0.605, 0.203)	0.305
Not Defined	(2)	-0.244 (-2.382, 1.894)	0.384
Level of Intervention			
Point-of-Use	(16)	-0.216 (-0.620, 0.189)	0.273
Community	(2)	-0.127 (-1.747, 1.492)	0.500
Effectiveness★			
Protective	(15)	-0.060 (-0.432, 0.312)	0.735
Non-Protective	(3)	-0.937 (-2.105, 0.231)	0.075
Recall Period *			
7-Day	(15)	-0.093 (-0.472, 0.286)	0.607
3-Day	(2)	-1.119 (-5.550, 3.312)	0.192
Group Equivalence °		-0.299 (-0.747, 0.148)	0.169
Equivalent	(12)	-0.299 (-0.747, 0.148)	0.169
Non-Equivalent	(3)	-0.256 (-0.811, 0.299)	0.186

* Denotes study left out of analysis due to utilization of 14-day recall
° Denotes 3 studies left out of analysis due to lack of data
★ Effectiveness as stated by study authors

In the second grouping of stratified analyses, continuous variables of the follow-up or observation period of each of the 18 analyzed studies were stratified for a deeper understanding of potential effects on the control group. With this initial descriptive study, it was determined that the most logical manner by which to divide studies amongst a range of a continuous variable would be the utilization of an arithmetic mean. Each of these mean values and therefore the description of the upper and lower groupings can be found in Table 8. These analyses, too, yielded statistically insignificant, yet suggestive effects on control group disease in terms of log-mean reported differences across the intervention. All analyzed relationships reported a negative difference from the first time measurement point (first measure) to the last time measurement point.

Table 8. Results of Stratified Paired T-Testing By Low and High Levels of Continuous Variable Characteristic

Characteristic	No. Studies High, Low	High Mean log-difference, (95% CI)	P-Value, High	Low Mean log-difference, (95% CI)	P-Value, Low
Length of Follow-up Period (μ=32.11)	7, 11	-0.347 (-0.874, 0.181)	0.159	-0.117 (-0.658, 0.425)	0.642
Total Number of Follow-up Visits (μ=25.35)*	6, 11	-0.226 (-0.572, 0.121)	0.156	-0.207 (-0.812, 0.398)	0.463
Frequency of Follow-up Visits (μ=3.41)*	9, 8	-0.116 (-0.619, 0.387)	0.609	-0.323 (-1.034, 0.387)	0.317

* Denotes one study left out of analysis due to lack of data

The final cluster of stratified analyses of intervention effects on control group disease focused on aspects that most closely associated with the type of intervention. Of the 18 water quality interventions—chlorination, filtration, flocculation, hygiene, solar disinfection (SODIS) and combination interventions; none was associated with a statistically meaningful mean log-difference in the control group. No provision of safe storage as part of the water quality intervention was one of the more borderline insignificant t test results in this study (p=0.077). An important observation is that all of these insignificant findings have a single directionality in common—the negative sign. It is important to note, however, that we are 95% confident however, that the true mean change in self-reported diarrheal rates across all strata includes the null hypothesis of 1, or no change. These mean log-differences are reported in Table 9.

Table 9. Table 8. Results of Stratified Paired T-Testing By Intervention Type

Type of Intervention	No. Studies (% total)	Mean log-difference, (95% CI)	P-Value
Chlorination			
Chlorination	7 (38.9%)	-0.260 (-1.088, 0.569)	0.473
No Chlorination	11 (61.1%)	-0.172 (-0.587, 0.244)	0.378
Filtration			
Filtration	7 (38.9%)	-0.141 (-0.859, 0.577)	0.647
No Filtration	11 (61.1%)	-0.247 (-0.723, 0.229)	0.274
Flocculation			
Flocculation	2 (11.1%)	-0.093 (-4.143, 3.956)	0.818
No Flocculation	16 (88.9%)	-0.220 (-0.620, 0.180)	0.259
Hygiene			
Hygiene	4 (22.2%)	-0.107 (-1.313, 1.528)	0.825
No Hygiene	14 (77.8%)	-0.296 (-0.681, 0.090)	0.121
Safe Storage			
Safe Storage	5 (27.8%)	-0.193 (-0.729, 1.115)	0.593
No Safe Storage	13 (72.2%)	-0.359 (-0.764, 0.045)	0.077
SODIS			
SODIS	2 (11.1%)	-0.324 (-3.480, 2.832)	0.416
No SODIS	16 (88.9%)	-0.191 (-0.593, 0.210)	0.326
Combination			
Combination	8 (44.4%)	-0.012 (-0.550, 0.526)	0.959
Single	10 (55.6%)	-0.361 (-0.907, 0.185)	0.169

DISCUSSION

Results

The paired t testing results show that there are no significant effects observed on reported rates of diarrhea among control groups comparing pre- and post-intervention time points within various water quality intervention types, designs and settings, among other characteristics. Confidence intervals for each of the three measurement point paired observation t-tests as well as the stratified analyses cross the null value of zero and p-values are insignificant at the alpha=0.05 level. Therefore, we cannot reject the null hypothesis of no significant difference in the mean change of reported diarrhea among control group members of water quality interventions.

In total, we were able to describe and analyze nearly 40 different characteristics within each of the 18 interventions included in this analysis sample set. We were successful in identifying water quality intervention trials that have self-report of diarrhea as an outcome and in collecting supplemental data on these studies for subsequent analysis. Data obtained included baseline (or, in the absence thereof, first surveillance point) reported diarrhea and endline or last surveillance point reported diarrhea, which was analyzed using paired t-testing to determine the magnitude and significance of observed intervention effects on these rates among control group members. Across the 34 different measures and stratified characteristics upon which we analyzed, 33 (97.1%) were negative in direction, although overall p-values ranged from 0.075 to 0.959.

The objective of this study was to determine whether and under what circumstances there are changes in reported diarrhea among control group members in water quality intervention trials. Several competing or potentially synergistic hypotheses have been postulated to explain the numerous instances in which reported diarrheal disease rates decrease among control group members when they are theoretically not receiving a direct intervention impact that may be beneficial.

Intervention Effect Spillover

One hypothesis is that of a spillover effect, which may have to do with overall intervention impact or effectiveness. The idea here is that an intervention is decreasing diarrheal disease levels in the intervention group such that the beneficial effects spread out from or spillover from this population due to lower environmentally associated disease levels. If an intervention is effective, then it may have a greater probability of effect on control group members, especially if they live in the immediate vicinity of the intervention group members. Although this was an initial, crude analysis, there was no significant relationship reported between the protective and non-protective interventions in terms of reported changes in levels of control group diarrhea [(protective intervention log-percent change: -0.060 (-0.432, 0.312); non-protective intervention log-percent change: -0.937 (-2.105, 0.231)].

Misclassification of the Control Group

Another hypothesis is that the control groups are being misclassified as such because they may be able to receive direct intervention effects during the observational period. An example may be drawn between household-level interventions (e.g., point-of-use water treatment) where one household assigned to a control groups could share an intervention assigned to a neighbor) and those that act on the community or shared water supply level. If the intervention happens to be community level, however, the potential for control group members to receive treatment-group water is theoretically increased. An assessment of the stratified data yielded no significant differences between POU [log-percentage difference -0.216 (-0.620, 0.189)] and community level [log-percentage difference -0.127 (-1.747, 1.492)].

The "Bugger-Off" Effect

A third hypothesis is very much behaviorally based: the idea that control group members who have had their patience run thin with a study that just won't end or study workers who visit too frequently or too many times end up reporting lower or at least less accurate rates of diarrhea to please data collectors. The stratified analyses of some of the control group characteristics observed in this study of water quality interventions were performed to judge their ability to explain a phenomenon that has been termed the "bugger-off" effect by some researchers (Clasen TF, 2013, Cairncross S 2013 personal communication, electronic mail). Of focus within this context were the continuous variables characterizing the follow-up period. The total number of follow-up weeks, the total number of data collection visits during this period and the frequency of follow-up data collection visits were three factors that could potentially play a role in this observed effect. The hypothesis here is that, the longer in duration or higher in frequency the follow up period or follow-up visits, respectively, the more likely it is that control group members will become exasperated by the repeated requests or extended duration of the study. Although this effect was not detected in this initial, descriptive study, an expansion of the literature pool upon which this question is based may yield a threshold period or frequency with which we begin to see a decrease in reported outcome data quality and/or response rate. Of the overall insignificant findings, the upper threshold of total duration of follow up (in weeks) and the total number of follow-up visits seemed to have a plausible significance within the purview of a future, expanded study ($p=0.16$, $p=0.16$, respectively).

Although Insignificant, an Observed Trend

Although there were no significant single relationships between water quality intervention characteristics and changes in reported diarrhea among control group members, the results taken as a whole suggest an interesting trend and potential relationship for future investigation. An understanding of the limitations of causal inference in this study are necessary, yet pervasive trends are noticeable even at relatively small sample sizes such as with this analysis of water quality interventions.

Of the three general paired t-test analyses we performed involving a comparison of baseline to last surveillance point, first surveillance point to last surveillance point, and a mixed first measure of baseline and first surveillance point to last surveillance point, all three were negative, yet insignificant relationships. Furthermore, over 97% of the 34 stratified analyses (33/34) performed by this study produced a negative (yet insignificant) effect on rates of reported diarrhea among control groups. The only measure that was determined to have a positive effect was observed in the non-RCT design water quality intervention trials (p=0.53). This spaghetti plot of the individual trends in reported diarrhea among all 18 studies provides a rough appraisal of the changes reported by the control group over the intervention period (figure 4). In order to show this relationship, all first measures gathered from supplementary data requests (baseline and first surveillance point) were utilized as a pre-intervention first measure as compared to the last surveillance measure.

An important realization is the context within which we are seeing this general trend of less reported disease in control groups over the intervention period—the study sample is quite small (n=18). A clear yet small negative trend can be observed with even a cursory glance, even though each study is utilizing different methodology for measuring diarrheal rates in its respective study population. Taking this idea further, Figure 5 shows the general trend in reported diarrhea rates among control group members even more clearly. Of the 18 intervention trials analyzed, 5 show an increase in diarrheal rates (27.8%) while 12 (66.7%) recorded a decrease in reported diarrheal disease among control members.

Methodological Strengths

The design of this study was the most appropriate for paired t testing. The independent observations were organized in pairs with a definite relationship within control groups, and the number of points in each dataset was the same. The design of the paired t-testing analysis allows for each control group to serve as its own control by only comparing the changes experienced in reported diarrhea by the same control group over the intervention period. This allows for good control of individual differences and variation between control groups, minimizing random error. To standardize this measure of effect and account for differences in modes and calculations of measurement of diarrheal rates, a percentage change from baseline (or first surveillance measure) to endline (last surveillance measure) was calculated. The mean and standard error of the differences within control groups over time is then calculated. The mean is then divided by the standard error of the mean to yield the test statistic, *t,* which is t-distributed with N-1 degrees of freedom.

Methodological Strengths and Advantages

Among the important advantages for this analysis was the fairly straightforward methodology behind sourcing all potential water and sanitation intervention trials from a recent WHO meta-analysis of such studies in low- and middle-income countries. Direct advantages included the fact that there was no need for a priori inclusion/exclusion criteria on the author of this study, but rather each included study had already "passed" an extensive round of inclusion testing. Also avoided were extensive and time-consuming database searches or significant time spent screening the literature as the group or 'universe' of studies was already decided. Extensive selection criteria and screening procedures were implemented in the inclusion of each study design, intervention type and setting amongst other characteristics. The author was able to proceed to the analysis of differences in reported diarrhea among control group members very quickly as there was no time spent on selection of the studies, only initial screenings for the specific purposes of the study of interest.

An important component of this study's strength is sourced from the nature of this novel research question. An advantage to this analysis is that it can be feasibly performed with a relatively small sample size (de Winter et al. 2013). Finally, the reduction in random error, or statistical "noise" allows us a greater likelihood of detecting any significant differences, if they do exist over independent different testing methods. Furthermore, the Wolf et al. 2014 paper was published in early 2014 and as such is the most current, comprehensive review of water and sanitation intervention literature available. Although this is most certainly an initial, descriptive analysis of water quality literature, it is nonetheless forging the first steps towards a more comprehensive future study of the effects on the control group in a water and sanitation intervention and research context.

Limitations

A Limited Universe

This study was vulnerable to a number of overarching limitations including combining measures of study effect and potential for a mixture of biases. This analysis was limited to analyzing relationships within and among the universe of intervention trials analyzed by the Wolf et al. 2014 WHO meta-analysis, a decision with distinct disadvantages. Several important limitations are connected to the sole inclusion of the Wolf et al 2014 meta-analysis database. The exclusive use of this database limited the universe of potential water quality interventions from which to request data, which severely limited data collection given the relatively low response rate and time constraints.

Limited to Supplementary Data Request Responses

Data for this study were obtained upon solicitation and subsequent receipt from study authors, making response to requests for these data criteria for study inclusion, by itself a potential selection bias. The difficulties experienced with communicating with researchers worldwide and very low response rates made the obtainment of supplementary data even more difficult, given this relatively small initial sample size of 72 intervention trials. Of the 47 studies for which the author requested supplementary data, about 35 responded in some form, eventually yielding only 18 completed supplementary data tables necessary for intervention control group analysis from 11 responding authors. Many of the requested data from older studies (studies older than 20 years) in the initial grouping of 47 screened studies were not included in this analysis due to the data being misplaced or destroyed, yielding additional potential or selection bias.

One of the requirements for inclusion within the WHO Meta-analysis was that the study had to be published. It is a known yet problematic factor in all of research as well as in water and sanitation literature that it is more difficult to publish studies without attaining a significant protective effect for the intervention group. Hence, when reviews fail to solicit grey literature for review, it excludes a differential percentage of non-significant trials, yielding biased estimates of effect for particular interventions. Along with grey literature, sanitation literature was, perhaps, underrepresented due to the inclusion criteria for the meta-analysis. Of the 72 total studies within the review, only 11 of them were primarily sanitation-related. Of the 18 studies included in this review, 4 involved hygiene education as a component of intervention, but none had an intervention focused on improving access to adequate sanitation.

Difficult Interpretation

The low numbers of studies in this analysis could have also led to potentially incorrect assumptions about the distribution of the data. With such a small sample, the author concedes that it is possible that even a sample of comparable size from a perfectly normal distribution could appear non-normal in a descriptive analysis. It is therefore possible that the log-transformation performed on these data was unnecessary, which would be unfortunate because it creates a more difficult interpretation of the study results reported in log-percentage change. After careful review of the associated univariate descriptive statistics, however, the log-transformation was deemed necessary to fulfill an important a priori assumption of normality for the paired observation sample distribution in paired t testing.

The treatment of each control group as an "individual" may be convenient for paired t-testing analyses, but this concept allows for errors to infiltrate measurements in the form of individual differences within each group. Unfortunately, the very methodology that made this study possible in a restricted time frame also did not allow for the collection or solicitation of individual-level data on intervention effects. Since not everyone within each control group will act, report diarrhea or react to the intervention (or theoretical lack thereof in this case), random error is produced.

Additional Potential Methodological Weaknesses

The analytical methodology required for this study created multiple levels of inherent error and potentially magnified biases. Effect estimates—the primary measure for analysis in this study—were calculated as a percent change to avoid dealing with some of the heterogeneity involved in outcomes of different interventions, which are themselves heterogeneous. Due to the nature of water and sanitation interventions, many of these studies are of lower methodological quality (Waddington et al. 2009; Clasen et al. 2010; Cairncross et al. 2010), which may impact results from their analysis. In fact, bias has been recognized as a near-inseparable characteristic of water and sanitation interventions, as the only blinded studies that minimized bias to date have not yielded significant effects (Clasen et al. 2006; Schmidt and Cairncross 2009).

The acceptability of some point-of-use interventions employed in many of the analyzed sample of trials has been questioned (Boisson et al. 2009) and is a potential concern for analysis of reported effects by control groups due to a potential selection bias of the sample population. If a particular intervention is not acceptable to a majority of the population in a particular area, intervention uptake may be characterized by lower enrollment rates. Although potentially vulnerable to significant biases, this initial descriptive study of effects on reported diarrhea among control group members was able to produce a systematically response to a majority of inherent and acquired biases.

The Inherent Issues of Diarrhea

A discussion of limitations for this study must first begin with the primary data of interest: the outcome measure and calculated percentage change in reported diarrhea among control group members. In some instances of low overall baseline or first surveillance diarrheal rates, very small changes in this rate led to large percentage changes over the course of the intervention. Since it was a percentage change, even a small change on a small initial rate may significantly skew the outcome measure. In addition, low levels of diarrhea may be harder to detect with accuracy due to lack of cases, leading to further potential bias. The impetus behind a percentage change calculation was the fact that different measuring methods were utilized to obtain still different measures of reported diarrhea, and all of these data needed to be standardized for aggregation and analysis.

Defining and Measuring Diarrhea

When stratifying on whether or not researchers defined diarrhea according to the standard WHO definition of three or more loose stools in a 24-hour period, (CITE), we found 16 studies followed this guideline while two studies left the interpretation of diarrhea up to survey respondents (Reller et al. 2003, Graf et al. 2010). Furthermore, there was disagreement with the definition of a "new" case of diarrhea. For example, the Luby et al. 2004 study in Pakistan utilized the standard WHO definition but went further in defining when a child was at risk for diarrhea. The study determined that children were only at risk for a new episode of diarrhea if they had reported no diarrhea in the preceding week (Luby et al 2004). Maeusezahl et al. 2009 defined a new episode of diarrhea as the occurrence of diarrhea after a period of 3 days symptom free. Sobsey et al. 2003, on the other hand utilizes the WHO definition but the caveat that a new episode of diarrhea be defined as one that had onset within the preceding week. As is evident, these varying definitions and applications of the standard definition of diarrhea opened the possibility for additional biases into this analysis.

Diarrhea Recall Period

Stratification was performed on diarrhea recall because the effect of this factor on self-reported diarrhea is still unclear. Although it has been shown by epidemiological study that recall of up to two weeks could be accurate (Blum and Feachem 1983; Graf et al. 2010), it has alternatively been shown that recall of as little as 7-days can lead to reporting errors (Alam et al. 1989) and recall beyond 2-days can introduce significant errors (Boerma et al. 1991). Still, other studies have shown that collecting data on both shorter periods and longer periods of recall—perhaps 2-3 day and two week recall—to be effective at estimating the true burden of disease (ICDDRB 2013).

Within the 18 analyzed studies in this analysis, study authors usually explicitly stated recall as 3-day or the more frequently used 7-day recall, yet there was one study with a 14-day recall (Graf et al 2010). There was also one study (Luby et al 2004) where the measure of diarrhea was not a 7-day recall but was determining whether diarrhea was experienced in the previous week or not (a week with or without diarrhea). For the purposes of this study analysis, this was deemed to be a 7-day recall. Although the specific magnitude or direction of the effect of varying lengths of diarrhea recall for these studies is not known, it is an important to understand these differences when combining analyses. Analysis of this differential in recall periods did not show a statistically significant relationship among the sample of intervention trials. Whereas the 7-day strata (n=15) reported a -0.093 [(-0.472, 0.286); p=0.607] log-percentage change, the interventions utilizing a 3-day recall (n=2) recorded a -1.119 [(-5.550, 3.312); p=0.192] log-percentage change in diarrheal rates reported among control group members.

Combining Measures of Effect

Another critical decision made in the performance of this analysis was in deciding to combine different estimates of incidence and prevalence of diarrhea and studies of different follow-up length to gain a broader perspective on changing rates of self-reported diarrhea within available water quality intervention literature. Due to difficulties in obtaining a sufficient amount of supplemental data for analysis to be able to make an attempt to answer the research question, it was necessary to combine all available measures of effect.

Among the prevalence estimates of self-reported diarrhea within included studies (n=10) was “longitudinal prevalence” (defined as the number of days with diarrhea divided by the total number of days of observation), “period prevalence” and just “prevalence.” Longitudinal prevalence of diarrhea was one of the most popular measures among the group of interventions in this analysis (n=5) is defined as the proportion of total observed person-time with disease outcome in individuals. It is a diarrheal morbidity measure that has been shown to be strongly correlated with adverse outcomes such as the risk of mortality in children under five years of age (Morris et al. 1996; Brown et al. 2008).

Longitudinal prevalence may also be more practical than incidence measures in terms of analysis because difficult-to-collect case frequency and duration are not necessary (Baqui et al. 1991).

Among the incidence estimates of self-reported diarrhea within included studies (n=8), was "incidence per child year", "incidence, episodes per 100 person weeks" and just "incidence". The author felt combining various calculations for incidence and prevalence was necessary as the study was in need of sufficient material to answer the question of whether control groups experienced less diarrhea over the course of water quality interventions. The current thought is that intermittent monitoring of diarrhea on a monthly basis to estimate longitudinal prevalence is effective due to its efficiency over more logistically intensive monitoring while yielding statistically similar measures (Schmidt et al. 2007). Study results reported here are not able to point to a more or less effective frequency of measure due to the small study size and insignificant findings of differential effect.

Combining Data from Different Study Designs and Types

Although there were several groups of study design (RCT vs. Non-RCT) and type (Point-of-use level v. community level), these interventions had to be combined for analysis and associated data could still be quite different from one another within strata. The inclusion of non-randomized controlled trials was necessary at this time in the field of water and sanitation. Although it is strongly recommended that review authors refrain from combining data from non-randomized and randomized controlled trial designs (Cochrane

2014), this author felt it necessary to include all studies included by the previous Wolf et al. 2014 meta-analysis of water and sanitation interventions to maintain continuity in the methodology of this analysis.

The community-level interventions in this study (n=2) consisted of piped-water systems that went through a variety of treatments before piping. Since it was unclear as to what portion of households received piped water from each type of water treatment, they were lumped together as methodologically equal. For the purposes of this study, follow up weeks were defined as the total number of weeks of the intervention and baseline reporting periods, regardless of number of data collection visits. This time point ranged from the initiation of baseline data collection to the last data collection visit.

Blinding Status

The blinded status of intervention design is yet another characteristics of potential importance to consider for this analysis of control group reported diarrhea. Water and sanitation interventions are often structured in such a manner where blinding is often not possible, and may not be desirable due to decreased compliance (Hartinger *et al*. 2011). The significance of stratification on blinding status was borne in the belief that a blinded study design would yield differential effects on control groups when compared to non-blinded studies due to a minimization of biases. From our analysis, we found yet another insignificant relationship. For blinded studies, the mean difference in control group reported diarrhea was -0.024 (-0.216, 0.168), or a drop of 2. 4% over the intervention period (p=0.36). There were only three blinded studies within the Wolf et al. 2014 meta-analysis (Kirchhoff et al. 1985; Jain et al. 2010; Boisson et al. 2013) and it was felt in that study that the number was insufficient with which to define potential biases. Within this study, only two of these blinded trials were included due to lack of supplementary data. Hence, it may be unsurprising to learn of insignificant differential effects due to blinding status or non-blinding status within this group of 18 intervention trials. No blinded interventions of household water treatment have yielded clear evidence of positive effects on health (Clasen 2007).

Standardizing Length of Follow-Up

For the calculation of follow up for studies stating length of follow-up in months, coded "followupweeks" in the analysis, the number of months was multiplied by the average number of weeks in a month over the course of the year to obtain the number of weeks utilized for analysis. The follow-up length of some of these studies had to be averaged since the exact length of study duration in weeks could not be ascertained from the literature. The average value of weeks used for one month was 365.25/12 = 30.4375 days per month, and 30.4375/7 = 4.348 weeks per month. For instance, if a study reported a follow up duration of 5 months, it would be calculated to be 4.348*5 = 21.741 weeks for the purposes of this study.

Within the group of 18 analyzed studies, there were two primary studies of concern for this analysis. In the Sobsey et al 2003 study, there were two study sites with interventions of differing length (Bolivia was 6 months and Bangladesh was eight months). For this study, in lieu of additional information the author took the average length of time between the two studies for the full study follow-up period (7 months). In Luby et al. 2004, study group data were combined by the author in the supplementary data report form, so intervention durations were combined and an arithmetic mean calculated. For instance, within the Luby et al 2004 study, there was one group of neighborhoods observed for 5 months and another group for 6 months. The group sizes were comparable in terms of neighborhoods for both groups.

Frequency of visit

When studies did not report specific numbers of follow-up visits to their intervention and control households (Graf et al 2010, Luby et al 2004), "weekly" was assumed to be once weekly an the number of follow-up weeks was divided by 4.3 in order to equal the average number of visits per month to standardize the measure. In some included studies within this analysis, this measure and calculation was unclear. For the Luby et al 2004 study unclear, so a decision on a summary measure had to be made, ending up with 37 weeks of follow up, visits 2x per week on average, so 74 visits total calculated. These 74 visits were made over (37 weeks * 4.348 weeks/month) months, so about 8.7 visits per month, on average. The general formula used to calculate the frequency of data collection visits was # total visits / (#weeks/#weeks per month (4.348)) = number of visits per month.

Effectiveness

The effectiveness (protective effect) of each intervention was evaluated based off of the study author's reported results. Quite simply, if an intervention was reported to have a statistically significant protective effect on its subject population, it was deemed to be effective for this study analysis. For more complicated or mixed studies, if a particular intervention had multiple intervention arms and/or combined intervention types and reported a mixture of overall effectiveness the overall study was deemed to be effective for the purposes of analysis if at least one intervention treatment was protective.

Attrition Bias

According to the Cochrane collaboration, attrition bias refers to the systematic differences between groups in withdrawals from a study (Cochrane 2014). Although this was not confirmed with a quantitative assessment, there were instances of household dropping out of control and intervention groups in each of these studies for various reasons. A future analysis would do well to document any potential for a differential attrition among these groups as a whole to determine if, for instance, greater numbers of control group members were dropping out of water quality interventions than the intervention group members because they weren't getting any better quality water. Another important factor to look at would be whether study length of follow-up, total number of data collection visits or frequency of data collection visits had any relationship to play in subject attrition over time.

Hawthorne (Observer) Bias

The impact of observer bias in analyses of control groups must be another factor of concern when performing such combined analyses. Study authors have noted that households within intervention groups may overstate their compliance with a particular water quality intervention (Graf et al. 2008; Graf et al. 2010). This effect may also hold importance for analyses on control groups (Hróbjartsson et al 2012). If there are frequent or even just repeated observations on control groups, they are more likely to, over time, report survey responses (and associated outcomes of interest) to please the survey staff. Although some studies have taken steps to minimize this potential bias, it is still a potentially pervasive and easily prominent force in self-reported outcomes.

Characteristics Not Analyzed By This Study

Age Groups

Although not analyzed in this study, each intervention was composed of a study population with different age-groupings, which could have potentially had an effect on the self-reported episodes of diarrhea. One hypothesis would be that a differential reporting rate would be experienced between households where parents or caretakers are reporting for children and households where the children and adults are able to report diarrhea themselves.

Quality Assessment of Interventions

Although there was a quality score computed for each study by the WHO meta-regression from which these interventions were sourced, it was decided they should not be an analyzed component of analysis for this initial, descriptive study. The quality score calculated by Wolf et al 2014 was adapted from the Newcastle-Ottowa Quality Scoring Technique (Wells et al. undated).

Summary Effect Measures

Although this study looked at each intervention in terms of a reported effect, the analysis was stratification based on protective or non-protective effect of the intervention. Hence, the study was reported as "effective" or "not effective" according to the significance of protective outcomes as reported by study authors.

Environmental Factors

An important concept to understand would be the seasonality of diarrheal disease in many of the low- and middle-income settings described in each of the included interventions on water quality. If a study follow-up period covered only part of the year, it is very possible that reported rates of diarrhea for all study participants will change over that time due purely to seasonality and mix with intervention-related effects (Wang et al. 2010). The control group is especially susceptible to this sort of seasonal change in diarrhea rates if they are receiving little, if any, augmenting water quality treatment materials over the course of the intervention.

Conclusion & Recommendations

A Basis for Expanded Study

The current study is but an initial foray into the analysis of control group effects across a range of water quality interventions. Despite a lack of significant findings, we cannot rule out the possibility that this was due to insufficient power due to the comparative paucity of data that could be included in the analysis. Observations and the critical nature of this research question focus on the foundation of experimental research: the control-intervention group comparison and the basis for causal inference. Such basic premises as the control group approximating the counterfactual and the control and intervention groups differing significantly on only the intervention itself help to formulate the conditions under which experimental research can lead to new and more comprehensive understanding of factors of important public health significance. As such, we suggest a revisit and expansion of this analysis in the near future. Above all, more data should be included. This not only includes more studies, but also more data points within those studies. If the complete data sets from these studies can be procured, this will also allow investigation of individuals within the control group rather than the control group collectively, thus increasing study power. A more robust analysis of study quality and intervention effectiveness would allow for a more informed stratification of intra-intervention changes in control group self-reported diarrhea. This analysis is in need of expansion to additional intervention types, settings, target populations and research designs to ascertain a greater understanding of the potential effects on control groups.

As the deadline approaches for the Millennium Development goals, which call for a reduction in child mortality by two-thirds from 1990 to 2015, we come to realize how much progress has been made and how effective, meaningful controlled studies of water quality and sanitation play a fundamental role in accomplishing the work that is still left to do (UNICEF & WHO 2009). If we are to reach this goal, a greater the knowledge of intervention effect on all stakeholders is an essential component of a growing understanding of activities and treatments that can work to reduce the burden of diarrheal disease and lessen overall health impacts on those in low- and middle-income countries.

To our knowledge, this was the first study that looked specifically at changes in reported diarrhea among control group members in water quality interventions. The potential bearings of this question are undeniably important and potentially wide-ranging; casting scrutiny on the very essence of experimental research studies in water and sanitation. If there is an observed effect on control group members, the potential impacts of acquiring the knowledge as to how this may occur may prove integral in the next generation of sustainable global health infrastructure development.

REFERENCES

AFMC (2014) AFMC Primer on Population Health; Part 2 – Methods: Studying Health: Chapter 5 Assessing Evidence and Information. The Association of Faculties of Medicine of Canada Public Health Educators Network. Retrieved from: http://phprimer.afmc.ca/Part2-MethodsStudyingHealth/Chapter5AssessingEvidenceAndInformation/Experimentalorinterventionalstudies

Alam N., Henry FJ., Rahaman MM. (1989) Reporting errors in one-week diarrhea recall surveys: experience from a prospective study in rural Bangladesh. Int J Epidemiol. 1989 Sept;18(3):697-700. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/2807676

Altman DG., Bland JM. (2011) Research Methods & Reporting, Statistics Notes: How to obtain the P value from a confidence interval. *BMJ 2011; 343 doi:* http://dx.doi.org/10.1136/bmj.d2304 (Published 8 August 2011)

Arnold, B.F. *et. al*. (2010) Causal Inference Methods to Study Nonrandomized, Preexisting Development Interventions. PNAS 107:22605. Retrieved from: http://www.pnas.org/content/107/52/22605.full

Arnold BF et al. (2013) Optimal recall period for caregiver-reported illness in risk factor and intervention studies: a multicountry study. Am J Epidemiol. 2013 Feb 15;177(4):361-70. doi: 10.1093/aje/kws281. Epub 2013 Jan 29. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/23364878

Baqui AH, Black RE, Yunus M, Hoque AR, Chowdhury HR, Sack RB. (1991) Methodological issues in diarrheal disease epidemiology: definition of diarrheal episodes. Int J Epidemiol 20: 1057–1063.

Boisson S, Stevenson M, Shapiro L, Kumar V, Singh LP, Ward D & Clasen T (2013) Effect of household-based drinking water chlorination on diarrhoea among children under five in Orissa, India: A double-blind randomised placebo-controlled trial. *PLoS Medicine* **10**(8), e1001497.

Brown J, Sobsey MD, Loomis D. (2008) Local water quality filters reduce diarrheal disease in Cambodia: a randomized, controlled trial of the ceramic water purifier. Am J Trop Med Hyg 2008; 79: 394–400.

Brown J, Hien VH, McMahan L, Jenkins MW, Thie L,Liang K, Printy E & Sobsey MD (2013) Relative benefits of on-plot water supply over other "improved"sources in rural Vietnam. *Tropical Medicine & International Health* **18**(1), 65–74.

Cairncross S, Hunt C, Boisson S, Bostoen K, Curtis V, Fung I, Schmidt WP. (2010) Water, sanitation and hygiene for the prevention of diarrhoea. *Int. J. Epidemiol. (2010) 39 (suppl 1): i193-i205. doi: 10.1093/ije/dyq035.* Retrieved from: http://ije.oxfordjournals.org/content/39/suppl_1/i193.full

Campbell DT, Stanley JC. (1966) *Experimental and Quasi Experimental Designs.* Chicago, Ill: Rand McNally; 1966.

Centers for Disease Control and Prevention (2013) Diarrhea: Common Illness, Global Killer. Retrieved from: http://www.cdc.gov/healthywater/global/diarrhea-burden.html

Clasen T, Garcia Parra G, Boisson S, Collin S. (2005). Household-based ceramic water filters for the prevention of diarrhea: a randomized, controlled trial of a pilot program in Colombia. Am J Trop Med Hyg 2005; 73: 790–5.

Clasen TF, Brown J, Collin S, Suntura O, Cairncross S. (2004) Reducing diarrhea through the use of household-based ceramic water filters: a randomized, controlled trial in rural Bolivia. Am J Trop Med Hyg 2004; 70: 651–7.

Clasen TF, Brown J, Collin SM. (2006) Preventing diarrhoea with household ceramic water filters: assessment of a pilot project in Bolivia. Int J Environ Health Res 2006; 16: 231–9.

Clasen TF., Schmidt WP., Rabie T., Roberts I., Cairncross S. (2007) Interventions to improve water quality for preventing diarrhea: systematic review and meta-analysis. BMJ 2007. *334* doi: http://dx.doi.org/10.1136/bmj.39118.489931.BE

Clasen TF. (2013) Comments from Dr. Thomas Clasen on a Draft of the GiveWell Water Quality Report, November 15, 2013. Retrieved from: http://files.givewell.org/files/DWDA%202009/Interventions/Water/Water%20Purification%20Assessment/thomas-clasen-comments.pdf

Cochrane Handbook (2014) Chapter 13: Including non-randomized studies. Barnaby C Reeves, Jonathan J Deeks, Julian PT Higgins and George A Wells on behalf of the Cochrane Non-Randomised Studies Methods Group. Retrieved from http://handbook.cochrane.org/

de Heer HD., Koehly L., Pederson R., Morera O. (2011) Effectiveness and spillover of an after-school health promotion program for Hispanic elementary school children. Am J Public Health. 2011 Oct;101(10):1907-13. doi: 10.2105/AJPH.2011.300177. Epub 2011 Aug 18. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/21852659 de Winter JCF. (2013) Using the Student's t-test with extremely small sample sizes. Practical Assessment Research & Evaluation. Vol 18:10. Published Aug 2013. ISSN 1531-7714

Feikin DR., Audi A., Breiman RF et al. (2010) Evaluation of the optimal recall period for disease symptoms in home-based morbidity surveillance in rural and urban Kenya. Int J Epidemiol. Apr 2010; 39(2): 450–458. Published online Jan 20, 2010. doi: 10.1093/ije/dyp374. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2846445/

Fewtrell L, Kauffman RB, Kay D, Enanoria W, Haller L, Colford JM Jr. (2005) Water, Sanitation, and hygiene interventions to reduce diarrhea in less developed countries: a systematic review and meta-analysis. Lancet Infect Dis. 2005 Jan;5(1):42-52. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/15620560

Graf, J., S. Z. Togouet, N. Kemka, D. Niyitegeka, R. Meierhofer and J. G. Pieboji. Health gains from solar water disinfection (SODIS): Evaluation of a water quality intervention in Yaounde, Cameroon. J Water & Health 2010; 8(4): 779-796.

Hróbjartsson A., Thomsen AS., Emanuelsson F., Tendel B., Hilden J. (2012) Observer bias in randomized clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012; 344 doi: http://dx.doi.org/10.1136/bmj.e1119 (Published 27 February 2012)

Hsu, H. and Lachenbruch, P. A. 2008. Paired *t* Test. Wiley Encyclopedia of Clinical Trials. 1–3. Food and Drug Administration, Rockville, MD, USA. Published Online: 19 SEP 2008. Wiley Encyclopedia of Clinical Trials. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/9780471462422.

ICDDRB (2013) Improving estimates of diarrhea prevalence among Bangladeshi children from survey data. Health and Science Bulletin, Vol. 11 No. 3. September 2013. International Centre for Diarrheal Disease Research, Bangladesh.

Jain S, Sahanoon OK, Blanton E, Schmitz A, Wannemuehler KA, Hoekstra RM & Quick RE (2010) Sodium Dichloroisocyanurate Tablets for Routine Treatment of Household Drinking Water in Periurban Ghana: A Randomized Controlled Trial. *The American Journal of Tropical Medicine and Hygiene* **82**(1): 16–22.

Jepsen P, Johnsen SP, Gillman MW, Sørensen HT. (2004) Interpretation of observational studies. Heart. 2004;90:956–960. doi: 10.1136/hrt.2003.017269. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1768356/#!po=72.2222

Kaptchuk TJ. (2001) The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf?. J Clin Epi 54(2001) 541-549.

Lesaffre E., Verbeke G. (2005) Clinical Trials and Intervention Studies.Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd. Edited by Brian Everitt and David Howell. ISBN: 0-470-86080-4. Retrieved from: http://www.wiley.com/legacy/wileychi/eosbs/pdfs/bsa099.pdf

Mäusezahl D, Christen A, Pacheco GD, Tellez FA, Iriarte M, Zapata ME, Cevallos M, Hattendorf J, Cattaneo MD, Arnold B, Smith TA & Colford JM Jr (2009) Solar drinking water disinfection (SODIS) to reduce childhood diarrhoea in rural Bolivia: a cluster-randomized, controlled trial. *PLoS Medicine* **6**(8), e1000125.

Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, Amann M, Anderson HR, Andrews KG, Aryee M. (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**(9859), 2224–2260.

Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, Rudan I, Campbell H, Cibulskis R, Li M, Mathers C, Black RE. (2012) Child Health Epidemiology Reference Group of WHO and UNICEF. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet. 2012; 379(9832):2151-61. Retrieved: http://www.ncbi.nlm.nih.gov/pubmed/22579125

Luby SP, Agboatwalla M, Hoekstra RM, Rahbar MH, Billhimer W, Keswick BH. (2004) Delayed effectiveness of home-based interventions in reducing childhood diarrhea, Karachi, Pakistan. Am J Trop Med Hyg 2004; 71: 420–7.

Luby SP, Agboatwalla M, Painter J. (2006) Combining water quality treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Trop Med Int Health 2006; 11: 479–89.

Majuru, B., M. M. Mokoena, P. Jagals and P. R. Hunter. (2010) Health impact of small-community water supply reliability. Int J Hyg & Env Health 214(2): 162-166.

Morris SS, Cousens SN, Kirkwood BR, Arthur P, Ross DA. (1996) Is prevalence of diarrhea a better predictor of subsequent mortality and weight gain than diarrhea incidence? Am J Epidemiol 144: 582–588.

NIH (2014) Prevention Research at NIH: Controlled Intervention Trials. Office of Disease Prevention. Division of Program Coordination, Plannign and Strategic Initiatives. National Institutes of Health Bethesda, Maryland. Retrieved from: https://prevention.nih.gov/prevention-research/phases-prevention-research/controlled-intervention-trials

Northwestern University Basic Sciences (1997) Possible alternatives if your data violate the paired t-test assumptions. Prophet StatGuide. Northwestern University

Feinberg School of Medicine. Retrieved from: http://www.basic.northwestern.edu/statguidefiles/ttest_paired_alts.html

Null C., Kremer M., Miguel E., VanDusen E., Zwane A. (2009) Measuring Diarrhea: Quantifying Hawthorne Effects in Frequently-Collected Data. The Bill & Melinda Gates Foundation, Innovations for Poverty Action. Retrieved from: http://siteresources.worldbank.org/INTDEVIMPEVAINI/Resources/3998199-1286435433106/BWM_WB_DIME_20091124.pdf

Pappas PA., DePuy V. (2004). An overview of non-parametric tests in SAS: when, why, and how. *Proc. SouthEast SAS Users Group Conference* (SESUG 2004): paper TU04. Duke Clinical Research Institute, Durham, 1-5.

Quick, R. E., A. Kimura, A. Thevos, M. Tembo, I. Shamputa, L. Hutwagner and E. Mintz. (2002) Diarrhea prevention through household-level water disinfection and safe storage in Zambia. Am J Trop Med Hyg 2002; 66(5): 584-589.

Ramakrishnan R., Venkatarao T., Koya PK., Kamaraj P. (1998) Influence of recall period on estimates of diarrhea morbidity in infants in rural Tamil Nadu. Indian J Public Health. 1998 Jan-Mar;42(1):3-6. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/10389497

Reller, M. E., C. E. Mendoza, M. B. Lopez, M. Alvarez, R. M. Hoekstra, C. A. Olson, K. G. Baier, B. H. Keswick and S. P. Luby. (2003) A randomized controlled trial of household-based flocculant-disinfectant water quality treatment for diarrhea prevention in rural Guatemala. Am J Trop Med Hyg 2003; 69(4): 411-419.

Rothwell PM. (2005) Treating Individuals 1: External validity of randomized controlled trials: “To whom do the results of this trial apply?” Lancet 2005 365: 82-83. Retrieved from: http://apps.who.int/rhl/Lancet_365-9453.pdf

Rothwell PM. (2006) Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials 1: e9.10.1371/journal.pctr.0010009. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3943905/

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. (2012) Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess. 2012 Sep;16(35):1-82. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/22989478

Schmidt WP, Arnold BF, Boisson S, Genser B, Luby SP, Barreto ML, Clasen T, Cairncross S.. (2011) Epidemiological methods in diarrhea studies—an update. Int J Epidemiol. Dec 2011; 40(6): 1678–1692. Published online Nov 8, 2011. doi: 10.1093/ije/dyr152. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3235024/

Schmidt WP, Luby SP, Genser B, Barreto ML & Clasen T (2007) Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology 18, 537–543.

Shadish WR., Cook TD., Campbell DT. (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company. Retrieved from: http://post.queensu.ca/~hh11/assets/applets/Causal_infe rence_in_experimental_and_quasi-experimental_designs.pdf

Sibbald B., Roland M. (1998) Understanding controlled trials: Why are randomized controlled trials important? *BMJ 1998; 316: 201 (Published 17 January 1998) doi: http://dx.doi.org/10.1136/bmj.316.7126.201*

Sobsey, M. D., T. Handzel and L. Venczel (2003) Chlorination and safe storage of household water quality in developing countries to reduce waterborne disease. Water Sci & Tech 2003; 47(3): 221-228.

Stauber, C. E., E. R. Printy, F. A. McCarty, K. R. Liang and M. D. Sobsey (2003) Cluster randomized controlled trial of the plastic BioSand Water filter in Cambodia. Env Sci & Tech 2012; 46(2): 722-728.

Stauber, C. E., G. M. Ortiz, D. P. Loomis and M. D. Sobsey. (2009) A randomized controlled trial of the concrete biosand filter and its impact on diarrheal disease in Bonao, Dominican Republic. Am J Trop Med Hyg 2009; 80(2): 286-293.

Steckler A, McLeroy KR. (2008) The importance of external validity. *American Journal of Public Health*. 2008;98(1):9–10. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2156062/#!po=62.5000

Talley NJ., Weaver AL., Zinsmeister AR., Melton LJ. (1994) Self-reported diarrhea: what does it mean? Am J Gastroenterol. 1994 Aug;89(8):1160-4. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/8053428

Tiwari, S. S. K., W. P. Schmidt, J. Darby, Z. G. Kariuki and M. W. Jenkins. (2009) Intermittent slow sand filtration for preventing diarrhoea among children in Kenyan households using unimproved water sources: Randomized controlled trial. Trop Med Int Health 2009; 14(11): 1374-1382.

UNICEF & WHO (2009) Diarrhoea: Why children are still dying and what can be done. World Health Organization. Geneva, Switzerland.

Wang, X., Hunter PR. (2010) A Systematic Review and Meta-Analysis of the Association between Self-Reported Diarrheal Disease and Distance from Home to Water Source. Am J Trop Med Hyg. Sept 2010; 83(3): 582-584. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929055/

WHO (1999) Cancer Epidemiology: Principles and Methods; Chapter 7: Intervention Trials. International Agency for Research on Cancer. Edited by I. Dos Santos Silva. World Health Organization, Geneva, Switzerland. Retrieved from: http://w2.iarc.fr/en/publications/pdfs-online/epi/cancerepi/

WHO (2008) Water Quality Interventions to Prevent Diarrhea: Cost and Cost-Effectiveness. TF Clasen, Haller L. Public Health and the Environment, World Health Organization, Geneva, Switzerland.

WHO (2013) Diarrhoeal Disease. Fact Sheet No 330. April 2013. Retrieved from: http://www.who.int/mediacentre/factsheets/fs330/en/

WHO/UNICEF (2013) Progress on Sanitation and Drinking-Water, 2013 Update. Joint Monitoring Programme for Water Supply and Sanitation. Geneva, Switzerland. Retrieved from: http://www.wssinfo.org/fileadmin/ user_upload/resources/JMPreport2013.pdf

Wolf J., Prüss-Ustün A., Cumming O., Bartram J., Bonjour S., Cairncross S., Clasen TF., Colford JM, jr., Curtis V., De France J., Fewtrell L., Freeman MC., Gordon B., Hunter PR., Jeandron A., Johnston RB., Mäusezahl D., Mathers C., Niera M., Higgins JPT. (2014) Assessing the impact of drinking-water and sanitation on diarrhoeal disease in low- and middle-income settings: A systematic review and meta-regression. World Health Organization Geneva, Switzerland.

Zimmerman, D.W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education, 67, 55-68.

Zwane AP, Zinman J, Van Dusen E, Pariente W, Null C, Miguel E, Kremer M, Karlan DS, Hornbeck R, Giné X, Duflo E, Devoto F, Crepon B, Banerjee A. (2011) Being surveyed can change later behavior and related parameter estimates. Proc Natl Acad Sci U S A. 2011 Feb 1;108(5):1821-6.

APPENDIX

Supplementary Data Request Template:

								Baseline					Endline	
Author(s)	Title	Publication	Year of Intervention	Year Published	Country	Intervention Description	Measure of disease frequency (incidence, prevalanece, etc.(	Was there a baseline (pre-intervention) measurement of diarrhoea in your study population? (Yes/No)	If there was a baseline measurement, what was the baseline level in the **control** group?	If there was a baseline measurement, what was the baseline level in the **intervention** group?	If no baseline measurement, what was the level of disease in the **control** group at the first surveillance point?	If no baseline measurement, what was the level of disease in the **intervention** group at the first surveillance point?	What was the endline measurement of diarrhoea in the **control** group	What was the endline measurement of diarrhoea in the **intervention** group
Aiken, B. A., C. E. Stauber, G. M. Ortiz and M. D. Sobsey	An assessment of continued use and health impact of the concrete biosand filter in Bonao, Dominican Republic.	ropical Medicine & H	June to August 2007.	2011	Dominican Republic	POU, biosand filter (BSF).	Incidence							